# Channel Access Management for Massive Cellular IoT Applications

by

Hesham Moussa

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

**Examining Committee Membership**

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

| | |
|---|---|
| External Examiner: | Jelena Misic<br>Professor, Dept. of Computer science, Ryerson University |
| Supervisor: | Weihua Zhuang<br>Professor, Dept. of Electrical and Computer Engineering,<br>University of Waterloo |
| Internal Member: | Kshirasagar Naik<br>Professor, Dept. of Electrical and Computer Engineering,<br>University of Waterloo |
| Internal Member: | Zhou Wang<br>Professor, Dept. of Electrical and Computer Engineering,<br>University of Waterloo |
| Internal-External Member: | John Wen<br>Associate Professor, Dept. of Mechanical and<br>Mechatronics Engineering, University of Waterloo |

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

**Abstract**

As part of the steps taken towards improving the quality of life, many of everyday life activities as well as technological advancements are relying more and more on smart devices. In the future, it is expected that every electric device will be a smart device that can be connected to the internet. This gives rise to the new network paradigm known as the massive cellular IoT, where a large number of simple battery powered heterogeneous devices are collectively working for the betterment of humanity in all aspects. However, different from the traditional cellular based communication networks, IoT applications produce uplink-heavy data traffic that is composed of a large number of small data packets with different quality of service (QoS) requirements. These unique characteristics pose as a challenge to the current cellular channel access process and, hence, new and revolutionary access mechanisms are much needed. These access mechanisms need to be cost-effective, enable the support of massive number of devices, scalable, practical, and energy and radio resource efficient. Furthermore, due to the low computational capabilities of the devices, they cannot handle heavy networking intelligence and, thus, the designed channel access should be simple and light. Accordingly, in this research, we evaluate the suitability of the current channel access mechanism for massive applications and propose an energy efficient and resource preserving clustering and data aggregation solution. The proposed solution is tailored to the needs of future IoT applications.

First, we recognize that for many anticipated cellular IoT applications, providing energy efficient and delay-aware access is crucial. However, in cellular networks, before devices transmit their data, they use a contention-based association protocol, known as random access channel procedure (RACH), which introduces extensive access delays and energy wastage as the number of contending devices increases. Modeling the performance of the RACH protocol is a challenging task due to the complexity of uplink transmission that exhibits a wide range of interference components; nonetheless, it is an essential process that helps determine the applicability of cellular IoT communication paradigm and shed light on the main challenges. Consequently, we develop a novel mathematical framework based on stochastic geometry to evaluate the RACH protocol and identify its limitations in the context of cellular IoT applications with a massive number of devices. To do so, we study the traditional cellular association process and establish a mathematical model for its association success probability. The model accounts for device density, spatial characteristics of the network, power control employed, and mutual interference among the devices. Our analysis and results highlight the shortcomings of the RACH protocol and give insights into the potentials brought on by employing power control techniques.

Second, based on the analysis of the RACH procedure, we determine that, as the number of

devices increases, the contention over the limited network radio resources increases, leading to network congestion. Accordingly, to avoid network congestion while supporting a large number of devices, we propose to use node clustering and data aggregation. As the number of supported devices increases and their QoS requirements become vast, optimizing node clustering and data aggregation processes becomes critical to be able to handle the many trade-offs that arise among different network performance metrics. Furthermore, for cost effectiveness, we propose that the data aggregator nodes be cellular devices and thus it is desirable to keep the number of aggregators to minimum such that we avoid congesting the RACH channel, while maximizing the number of successfully supported devices. Consequently, to tackle these issues, we explore the possibility of combining data aggregation and non-orthogonal multiple access (NOMA) where we propose a novel two-hop NOMA-enabled network architecture. Concepts from queuing theory and stochastic geometry are jointly exploited to derive mathematical expressions for different network performance metrics such as coverage probability, two-hop access delay, and the number of served devices per transmission frame. The established models characterize relations among various network metrics, and hence facilitate the design of two-stage transmission architecture. Numerical results demonstrate that the proposed solution improves the overall access delay and energy efficiency as compared to traditional OMA-based clustered networks.

Last, we recognize that under the proposed two-hop network architecture, devices are subject to access point association decisions, i.e., to which access point a device associates plays a major role in determining the overall network performance and the perceived service by the devices. Accordingly, in the third part of the work, we consider the optimization of the two-hop network from the point of view of user association such that the number of QoS satisfied devices is maximized while minimizing the overall device energy consumption. We formulate the problem as a joint access point association, resources utilization, and energy efficient communication optimization problem that takes into account various networking factors such as the number of devices, number of data aggregators, number of available resource units, interference, transmission power limitation of the devices, aggregator transmission performance, and channel conditions. The objective is to show the usefulness of data aggregation and shed light on the importance of network design when the number of devices is massive. We propose a coalition game theory based algorithm, *PAUSE*, to transform the optimization problem into a simpler form that can be successfully solved in polynomial time. Different network scenarios are simulated to showcase the effectiveness of *PAUSE* and to draw observations on cost effective data aggregation enabled two-hop network design.

**Acknowledgements**

I would like to extend my deepest gratitude and sincerest appreciation to my supervisor, Professor Weihua Zhuang, for her continuous help, guidance, encouragement and support throughout my Ph.D. study and my personal life at University of Waterloo. Being an outstanding supervisor and extremely insightful, she guided me through my Ph.D. program and provided me with valuable comments that have always inspired me to do in-depth thinking on my research. She taught me how to be a good researcher and successful Professor. I learned from her to accept people's criticism in a positive manner and use the feedback to improve my work and my skills.

I would like to also thank Professor Zhou Wang, Professor Saggar Naik, Professor John Wen, and the external examiner, Professor Jelena Misic, from Ryerson University, for serving in my Ph.D. research committee. Their suggestions, valuable comments and insightful questions have aided in the significant improvement of the presented work in this thesis.

Thanks are also due to Professor Xuemin (Sherman) Shen for his great support who, along with Professor Zhuang, has set an exemplary model for the successful professor who I aspire to become one day. They both have always been a major source of inspiration. Their work ethics are unprecedented and set the bar for how every professor should be.

Additionally, over my past four years in the Ph.D program, I have gained acquaintance with many friends and colleagues from the Broadband Communication Research (BBCR) group with whom I have enjoyed every moment. They also helped me through the hardships. Without their presence, this work would not have been possible.

Finally, I can never thank enough my dear parents and brother who, although are thousands of miles away, have been with me in every step in my life and whose prayers have made this moment possible. Thank you all for everything.

**Dedication**

*Praise be to Allah*

*To my dear parents, Gamal and Mona, amazing brother Amr,*
*deceased brother Ahmad, and great mentor Weihua Zhuang*

# Table of Contents

# List of Tables

# List of Figures

# Acronyms

| | |
|---|---|
| **3GPP** | 3rd Generation Partnership Project |
| **ACB** | Access Class Baring |
| **AP** | Access Point |
| **APA** | Access Point Association |
| **BS** | Base Station |
| **CDF** | Cumulative Density Function |
| **CRE** | Cell Range Expansion |
| **DA** | Data Aggregator |
| **EAB** | Enhanced Access Baring |
| **FDMA** | Frequency Division Multiple Access |
| **FPC** | Fractional Power Control |
| **IEEE** | Institute Of Electrical And Electronics Engineers |
| **IETF** | Internet Engineering Task Force |
| **IoT** | Internet Of Things |
| **M2M** | Machine-to-Machine |
| **MAC** | Medium Access Control |
| **MMTC** | Massive Machine Type Communication |
| **MTC** | Machine Type Communication |
| **NOMA** | Non-orthogonal Multiple Access |
| **OMA** | Orthogonal Multiple Access |
| **PCP** | Point Cluster Process |
| **PDCCH** | Physical Downlink Control Channel |
| **PDF** | Probability Density Function |
| **PMF** | Probability Mass Function |
| **PPP** | Point Poisson Process |
| **PRACH** | Physical Random Access Channel |
| **PUCCH** | Physical Uplink Control Channel |
| **QoS** | Quality Of Service |

**RACH** Random Access Channel

**RAR** Random Access Response

**RAT** Radio Access Technology

**RB** Resource Block

**RSS** Received Signal Strength

**SCHN** Small Cell Heterogeneous Networks

**SIC** Successive Interference Cancellation

**SINR** Signal To Interference Plus Noise Ratio

**SIR** Signal To Interference Ratio

**TDMA** Time Division Multiple Access

**UA** User Association

**WSN** Wireless Sensor Networks

# Nomenclature

| | |
|---|---|
| $\mathcal{A}$ | Denote the cardinality of the set $\Phi_a$ |
| $\mathcal{B}$ | Denote the cardinality of the set $\Phi_b$ |
| $C_d$ | NOMA Device-DA coverage probability |
| $C_{tot}$ | NOMA Device-BS total coverage probability |
| $\mathcal{D}$ | Total number of devices in the network |
| $\mathcal{E}_a$ | Average energy consumption of a DA to transmit a single data packet |
| $\mathcal{E}_a^f$ | Fixed amount of energy consumed by a DA in the case of failed access attempt |
| $\mathcal{E}_d$ | Average energy consumption of a NOMA device to transmit a single data packet |
| $\mathcal{E}_d^f$ | Fixed amount of energy consumed by a NOMA device in the case of failed access attempt |
| $\mathcal{O}$ | Total computational complexity of the *PAUSE* algorithm in number of iterations |
| $\mathcal{P}$ | Partition of the devices in the network |
| $\mathbf{Q}_b$ | The collection of the overall transmission power assignment vectors for BS associated devices |
| $\mathbf{Q}_{d_i,a_j}$ | Vector indicating the overall transmission power assignments of the $i^{th}$ device from the $h^{th}$ DA on all $n$ RCs |
| $\mathbf{Q}_{d_i,b_j}$ | Vector indicating the overall transmission power assignments of the $i^{th}$ device from the $j^{th}$ BS on all $K$ RBs |
| $\mathcal{R}_a^k$ | Average achievable packet rate of a DA |
| $\mathcal{R}_{a_h}^l$ | Achievable data rate by the $h^{th}$ DA towards its serving BS on the $l^{th}$ RU |
| $\mathcal{R}_d^{sjn}$ | Average achievable packet rate of a NOMA device of rank $j$ |
| $\mathcal{R}_{min}$ | Minimum average rate requirement of a device |
| $R_{0i}^n$ | Distance between $i^{th}$ device located inside the Voronoi cell of $BS_0$ and the origin |
| $\mathbf{S}$ | RB assignment vectors for all BS-associated devices in the network. |
| $\mathbf{S}_{d_i,a_h}$ | Vector indicating the overall RB assignments of the $i^{th}$ device from the $h^{th}$ DA on all $n$ RCs |
| $\mathbf{S}_{d_i,b_j}$ | Vector indicating the overall RB assignments of the $i^{th}$ device from the $j^{th}$ BS on all $K$ RBs |
| $T_a$ | Length of aggregation frame |
| $T_f$ | Length of full transmission frame |
| $T_r$ | Length of relaying frame |
| $\mathcal{V}(C_j,\mathcal{P})$ | Gain of the $j^{th}$ coalition, $C_j$, given the partition $\mathcal{P}$ |

| | |
|---|---|
| $x_{d_i,b_j}$ | Binary indication variable marking the association status of the $i^{th}$ device with the $j^{th}$ BS |
| $X^n_{ji}$ | Distance between the $i^{th}$ device, associated with the $j^{th}$ BS (for $j > 0$), and the origin |
| $y_{d_i,a_h}$ | Binary indication variable marking the association status of the $i^{th}$ device with the $h^{th}$ DA |
| $Y^n_{ji}$ | Distance between the $i^{th}$ device, associated with the $j^{th}$ BS (for $j > 0$), and its serving BS |
| $\alpha$ | Path-loss exponent |
| $\gamma_a$ | Packet arrival rate at a DA |
| $\gamma_d$ | Packet arrival rate at a NOMA device |
| $\Delta t_{e2e}$ | Average end-to-end system delay |
| $\epsilon$ | Fractional power control factor |
| $\bar{\Theta}$ | Average transmission success probability of a device |
| $\Theta(\Xi^k_a)$ | DA transmission success probability |
| $\Theta_{as}$ | Unconditional association success probability of a device using the RACH protocol |
| $\Theta(\Xi^{ns}_{ij})$ | NOMA device transmission success probability given the device ranking $j$ |
| $\lambda_a$ | Density of DA per km |
| $\lambda_b$ | Density of BS per km |
| $\lambda_d$ | Density of active devices per km |
| $\lambda_n$ | Density of interfering devices per km |
| $\Lambda_a$ | Radius of the disk like coverage area of a DA |
| $\bar{\Lambda}_s$ | NOMA sub-channel scheduling probability |
| $\mu_a$ | Packet departure rate from a DA |
| $\mu^{sjn}_d$ | Packet departure rate from a NOMA device |
| $\Xi^k_a$ | SIR of the received transmission from a DA |
| $\Xi^n_{ji}$ | SINR of the $n^{th}$ preamble transmitted by the $i^{th}$ device and received by the $j^{th}$ |
| $\Xi^{ns}_{ij}$ | SIR of the received transmission from a NOMA device on the $n^{th}$ sub-channel with rank $j$ |
| $\rho$ | Receiver sensitivity of the devices |
| $\varrho$ | A design threshold parameter for terminating the feasibility check stage |
| $\tau_a$ | SINR threshold of DA |
| $\tau_b$ | SINR threshold of BS |
| $\Upsilon_a$ | Super set of associated devices to all DAs in the network coalition formation game based partitioning |
| $\Upsilon_{a_h}$ | The set of devices associated with DA $h$ coalition formation game based partitioning |
| $\Upsilon_b$ | Super set of associated devices to all BSs in the network after coalition formation game based partitioning |

| | |
|---|---|
| $\Upsilon_{b_j}$ | The set of devices associated with BS $j$ after coalition formation game based partitioning |
| $\Phi_a$ | Set of DA locations |
| $\Phi_b$ | Set of BS locations |
| $\Phi_d$ | Set of device locations |
| $\Psi_a^k$ | Location set of the DAs interfering on the $k^{th}$ uplink cellular sub-channel |
| $\Psi_{nj}$ | Locations set of intra-cell interfering device on the $n^{th}$ preamble and associated with the $j^{th}$ BS |
| $\Psi_n^h$ | Location set of the primary interfering NOMA devices on the $n^{th}$ NOMA sub-channel |
| $\Psi_n^l$ | Location set of the secondary interfering NOMA devices on the $n^{th}$ NOMA sub-channel |

# Chapter 1

# Introduction

## 1.1 Massive M2M communication

The Internet of Things (IoT) is expected to have a high impact on several aspects of everyday-life as it will help realize many intelligent applications such as environmental monitoring, smart cities, intelligent transportation systems, and e-health applications. Applications of IoT also extends to improving the efficiency of industrial production, automation, and robotics. Essentially, IoT is anticipated to be a major economical player that will revolutionize earth in all aspects. It is anticipated that there will be over 64 billion connected devices by the year 2025, many of which are simple battery operated machines [57]. This gives rise to a new low-cost communication paradigm known as machine-to-machine (M2M) communication (or machine type communication (MTC)) which has been regarded as a fundamental enabling technology for many IoT applications. M2M communication is a technology whereby a number of machines communicate autonomously and collaborate to achieve a certain objective with minimal to no human intervention [21]. M2M devices are referred to as machine type devices or devices for short. Massive M2M applications are of especial interest as they pose new networking challenges to be addressed.

Massive M2M communication is characterized by a large number of devices that produce small data packets in a periodic or event-driven manner. Many devices are battery operated low complexity and low mobility devices that are equipped with limited radio front-ends. Being quite different from the traditional data-hungry human based applications which virtually have access to infinite operational power, massive M2M communication has intrigued many research and standardization efforts in the recent years [8]. According to various standardization bodies, such as the $3^{rd}$ generation partnership group (3GPP) [101], the Institute of Electrical and Electronics Engineers (IEEE) and the Internet Engineering Task Force (IETF) [8], M2M communication will become an important source of traffic in future 5G networks. Out of the various 5G enabling technologies, densified multi-tier cellular networks are seen as the most well suited technology for

supporting the anticipated explosion in the number of connected devices, thanks to their wide coverage area, high device capacity, low device energy consumption and mobility support features [21]. Nonetheless, the major obstacles crippling their immediate suitability for massive M2M communications are the load imbalance caused by the underlying sub-optimal user association algorithm, cost inefficiency of network deployment, and the inefficient traditional cellular random access channel (RACH) procedure [21, 35, 70]. Therefore, new networking solutions are needed to efficiently support the predicted massive number of heterogeneous devices on future cellular networks without resorting to expensive infrastructure alterations [31].

## 1.2 Motivations and objectives

The random channel access of the traditional cellular is not suitable to handle the unique characteristics of M2M applications for multiple reasons [69]. First, most M2M applications are event-driven, leading to a huge amount of simultaneous access requests sufficient to congest the physical random access channel (PRACH) [35]. A congested PRACH results in sensible network access delays, and reduces network good-put. Second, for M2M applications to be economically viable, devices are expected to stay operational for a long time with minimum maintenance. A congested PRACH results in energy wastage as devices have to initiate the RACH procedure multiple times before associating successfully. Yet, devices are usually placed at inaccessible locations, making it difficult to do battery replacement or use other sophisticated energy harvesting mechanisms; thus, other means should be exploited. Third, by coexisting with other applications, such as human-to-human communication, M2M applications are considered a major threat that may lead to a significant decrease in the overall network quality of service (QoS) as the radio channel becomes congested [116, 126]. Last, in addition to performing poorly under massive access, the random access process induces excessive signaling overhead, making them inefficient for many M2M applications which are characterized by small-sized data packets sporadically transmitted by a massive number of devices [21, 70]. Consequently, to abide by the delay and low energy consumption requirements of massive M2M communications and avoid degraded QoS performance, alleviating congestion at the PRACH is seen as a key enabling solution [67, 70, 120].

One way of alleviating the congestion of the PRACH is by introducing data aggregation, where some of the devices are off-loaded through powerful data aggregation nodes. However, utilizing data aggregators (DAs) leads to a multi-tier network architecture which introduces multiple layers of complexity, although having many benefits such as enhancing energy conservation of the devices by bringing the base stations (BSs) closer to them, increasing the number

of supportable devices by densifying the number of serving access points (APs), and improving the overall network performance by maximizing resource utilization and relieving congestion via data offloading mechanisms. Different from single-tier networks, in multi-tier networks, devices are located in areas with multiple APs to choose from. How devices associate with an AP plays an important role in determining the overall network performance [83, 110]. Therefore, to achieve the anticipated potentials of multi-tiered networks, optimizing user association is crucial. Yet, in the case of data aggregation infused cellular networks, the aggregation nodes are considered to be cellular devices that share the radio resources with other devices at the BSs. Thus, compared to traditional optimization of user association, under the considered scenario, the problem suffers from extensive dependencies between possible routes available for data transmission. Accordingly, the problem is more complex and is in fact a multi-objective optimization problem, the objective of which is to maximize the number of supported devices at their QoS requirements, while minimizing their transmission power and improving the overall resource utilization, taking into account the dependency between the devices and the DAs when sharing the resources. Thus, the applicability of the available user association mechanisms is limited and results in a sub-optimal solution for the inherently intertwined objectives for the purpose of M2M applications.

Motivated by the potentials of future cellular networks for supporting massive M2M communication, in this PhD research, we investigate ways to overcome the above discussed shortcomings to enable cost-effective and efficient massive cellular M2M communication. We tackle the subject from three inter-related perspectives. First, under heavy loading conditions, we analyze the performance of the RACH procedure used by the devices to access the PRACH to identify its limitations, and propose possible ways to improve its performance at the minimum possible infrastructural alteration. Second, we study the potential advantages brought by data aggregation for the purpose of massive cellular M2M communications as a way of enhancing the RACH performance as well as increasing its efficiency. However, different from the literature, we propose to use non-orthogonal multiple access (NOMA) at the DA as a way of enhancing radio resource utility and increasing the number of supported devices. Last, under data aggregation infused cellular networks, we formulate and efficiently and optimally solve a multi-objective optimization problem to improve energy efficiency and resource utilization via proper user association and power control, taking into account the unique characteristics of massive cellular IoT applications. A detail discussion and the respective contributions of these perspectives are given in Sections 1.3, 1.4, and 1.5 respectively.

## 1.3 Random access channel procedure

With the proliferation in IoT applications, a large network that is made up of a massive number of heterogeneous devices spread across a large surface area will be formed. Such unconventional network brings many new networking challenges that have attracted attentions from many researchers in recent years. Understanding the performance of a system is a key to overcoming its shortcomings. In traditional cellular communication, for a device to transmit its data, it has to first go through an association process that is done over the PRACH. According to the $3^{rd}$ Generation Partnership Project (3GPP), the number of IoT devices in the coverage range of a typical base station is expected to exceed thirty thousand devices. Due to the random access nature of the association process, with a such large number of contending devices, a massive number of simultaneous access requests will be generated, resulting in congestion of the PRACH. A congested PRACH introduces sensible network access delays, increases device energy consumption, and reduces network good-put [70]. Therefore, to provide ubiquitous energy efficient and timely connectivity, improving network access is crucial [74]. Yet, the solution should be economically viable and comes at minimum infrastructural alteration [70, 120].

In current cellular communication systems, devices use a contention-based 4-way handshake procedure, RACH, to associate with their desired BSs over the PRACH. Thus, to understand how cellular networks behave under massive M2M communications, analyzing the performance of the RACH procedure is pivotal. In the first step of the RACH protocol, an active device, with data to transmit, randomly chooses and transmits a preamble to its desired BS. The 3GPP defines a fixed set of orthogonal preambles which is shared by all the devices in the network. Thus, as the number of contending devices increases, the possibility of more than one device in a cell choosing the same preamble simultaneously increases, leading to extensive access delays and elevated levels of energy consumption [40]. This problem is known as the PRACH overload problem and has been well studied in the literature. In fact, many of the proposed solutions have already been implemented in the current cellular access network for the purpose of alleviating congestion and enhancing system throughput [70, 120]. However, these solutions are yet to be optimized for massive access scenarios such as IoT applications.

Thus, as a step towards improving the performance of the cellular access network when supporting large-scale IoT applications, many researchers have studied the RACH protocol and modeled its performance under massive access setting [44, 70, 75, 99]. Most of these studies are based on the assumption that all colliding devices will be denied access and have to reattempt the RACH protocol in subsequent opportunities [70]. Under this assumption, the RACH has performance equivalent to that of the slotted ALOHA protocol. While this assumption is valid,

with the recent improvements in the computational abilities of BSs and the introduction of various multi-user detection techniques, it becomes possible for colliding devices to successfully associate if their preambles are received with the signal to interference plus noise ratio (SINR) above a certain threshold [45, 106]. Hence, with proper interference mitigation mechanisms, the RACH performance can be enhanced. Accordingly, to investigate the RACH performance under this new scenario, novel SINR models are needed to capture the added complexity due to multi-user detection.

A powerful tool that has been used by many to model the SINR of various complex networks is stochastic geometry [90]. Although these works lay the grounds for developing a mathematical model for the RACH performance, they cannot be directly applied as the RACH procedure exhibits a wide range of interference components that are not well studied in the literature. In this research, with the help of stochastic geometry, we analyze the performance of the RACH procedure via modeling the in-network interference components and deriving a mathematical model for the association success probability. We consider a large number of stationary battery powered and identical devices attempting data transmission simultaneously. devices use power control which enables them to compensate for the attenuation due to propagation loss. The model considers both inter-cell and intra-cell interference components as well as the spatial distribution of the network. In this regards, the main contributions of this part of the work in RACH performance analysis are summarized as follows:

- We present a novel mathematical framework for modeling the performance of the RACH protocol under large-scale IoT applications. We model the SINR of the preamble transmission and derive an expression for the instantaneous association success probability;
- The developed model is comprehensive as it accounts for various network parameters including device density, spatial spread of the network, channel characteristics, association policy, device power control, and the number of available preambles. The model highlights the effect of both intra-cell and inter-cell interference factors on the success probability;
- The accuracy of the analytic model is corroborated via computer simulations;
- The joint effect of fractional power control (FPC), device density, and the SINR threshold on the success probability is studied and the main trade-offs are highlighted.

## 1.4 NOMA-enabled cluster-assisted network

The overload problem, where the massive simultaneous access of devices congests the PRACH, has been studied extensively and many solutions, such as ACB [70], EAB [42], randomized back-

off schemes [121], prioritized random access [76], and many others [67, 69], have been proposed. However, based on our analysis of the RACH procedure, we see that the most promising solution to enable massive MTC (mMTC) and avoid overloading the access channel is to used node clustering and data aggregation [66, 87]. With clustering and data aggregation, devices do not contend directly on the available resources at the BS; instead, they connect to a middle layer of DAs which relay the data on their behalf in a two-hop fashion. By doing so, the number of devices contending for BS radio resources decreases, reducing the chances for congestion. Furthermore, devices consume less energy as the transmission distance is significantly shortened [48, 54, 84]. Furthermore, clustering can decrease the RACH overhead ratio by aggregating small data packets into larger ones before transmission, leading to improved resource utilization.

Data aggregation has been utilized in other networks such as WSNs [1]. In WSNs. data aggregation is often implemented by carefully localizing powerful data aggregation nodes at optimal locations. However, this is feasible for networks of small to medium size. In the case of massive MTC, the number of devices is very large. Hence, for data aggregation to be cost effective, the cost of DA deployment should be kept to minimum, which may prohibit network wide deployment of powerful aggregation nodes. Consequently, in recent research, other means of cost-effective data aggregations have been proposed. Some suggest electing an MTC device to become the cluster head of each cluster, while other propose to use nearby mobile phones as relays [2, 53, 87]. Other researchers have also looked into utilizing UAVs, such as aerial drones, to provide dynamic and optimized data aggregation and clustering when needed [54]. How to aggregate data has been a hot topic of research in recent years. However, how to cluster is out of the scope of this work. The objective is to minimize the number of deployed DAs such that the network is cost effective while maximizing the number of supported devices at the required QoS.

Given these potentials, recently, node clustering and data aggregation have been investigated as possible solutions for enabling energy efficient mMTC while maintaining acceptable network performance in terms of delay and throughput [36, 74]. For instance, in [66], to study the impact of clustering on the outage probability under different channel models, the clustering problem is formulated to answer three main questions: how many clusters to form, what should be the transmit power of the devices, and how the clustering decision depends on the networking environment. The authors conclude that, with the proper choice of the number of clusters to form, optimized network design can be achieved. In [48], Guo et al. consider a two-phase transmission system where the DAs take the responsibility of not only relaying data, but also handling resource scheduling among active devices. It is shown that by properly choosing the number of

aggregators, the number of successfully served devices can be enhanced. In [84], the clustering problem is formulated as an optimization problem in order to determine the number of clusters which results in the lowest energy consumption of the devices, with a hierarchical transmission system of two or more hops. In [54], energy efficiency of cellular networks for massive IoT applications is studied under a drone assisted scenario. In [53], different deployment strategies for the DAs are studied in terms of their impact on energy consumption and coverage probability. In [87], an $n$-CSMA medium access for in-cluster transmission is proposed, and its impact on the performance of clustered networks is studied, including the correlation between cluster size and overall network performance in terms of energy and throughput. In [65], a two-stage access for cellular IoT is considered, and queuing theory is used to analyze the delay performance as a function of the device density. All the existing works emphasize on the effectiveness of node clustering and data aggregation for enabling mMTC, while maintaining acceptable network performance. They also show that, when dealing with a massive number of devices, data aggregation becomes challenging and optimization of different design parameters is essential.

On the other hand, NOMA has recently gained attention as a potential medium sharing scheme to improve the spectral efficiency and to increase the number of supported devices in future 5G networks [108]. In NOMA, thanks to multi-user detection and interference mitigation techniques such as successive interference cancellation (SIC) [104], multiple devices are allowed to share the same radio channel simultaneously, leading to a higher user capacity and better performance than traditional OMA [119]. Thus, a NOMA-based data aggregation framework presents a promising solution to enable mMTC. Nevertheless, most of the existing works on data aggregation for mMTC communications have focused on optimizing the network performance with OMA based channel access [48, 84], and only a few have considered NOMA [7].

While these studies demonstrate the potential of NOMA as a method for supporting massive connectivity in future cellular networks, further research on the potentials of NOMA-enabled clustered operation is needed. Thus, in an effort to further explore the potentials of node clustering and data aggregation in the context of mMTC, in this part of the research, we aim at shedding some light on the impact of clustering on different network performance metrics such as number of supported devices, energy consumption, and delay. We propose a new two-hop NOMA enabled clustered transmission framework to support massive cellular IoT applications. As we are considering NOMA, all of the above network parameters depend on the the severity of the interference, which is a function of the number of devices sharing the same resource channel as well as the transmission power of the devices. For this part, we mainly consider delay and energy efficiency as the QoS metrics required by the devices. Accordingly, to be able to quantify

the network performance metrics, we characterize the in-network interference components and define the coverage probability in the system. We then utilize techniques from both queuing theory and stochastic geometry to achieve tractable analytical results for the end-to-end delay and energy consumption of the devices and DAs as functions of the defined coverage probability. Our proposed framework captures the unique characteristics of future IoT applications, including the small sized data packets generated by the devices, massive number of devices in the network, and limited radio resources. Thus, in this regards, the main contributions of this part of the work are summarized as follows:

- We present a novel NOMA-enabled two-hop network model for massive cellular IoT communications. We develop a general analytical framework to obtain approximating yet accurate mathematical models for the coverage probability, average number of served devices, overall average access delay, and average energy consumption of a device and a DA;
- Compared to the literature, our derived expressions are more comprehensive and thus can be used to study the impact of various design parameters (such as device density, DA density, device transmission power, and available radio resources) on delay, coverage probability, and energy efficiency performance of massive cellular IoT applications;
- The accuracy of the analytic models are corroborated via computer simulations. Compared with traditional OMA-based clustered networks, with the proper choice of the number of clusters, our proposed solution improves the overall network performance.

### 1.4.1 Load balancing with data aggregation

Under data aggregation enabled approach, the cellular network is overlaid with a layer of DAs which function as data collection nodes and relays. In these networks, devices have the choice of directly connecting with BSs or going through DAs to transmit their data in a two-hop fashion. Data packets from DA associated devices are first aggregated into larger data packets and then transmitted from the DA to the BS. Data aggregation has the advantage of providing energy efficient communication as the DAs are often at a closer proximity to the devices than the cellular BSs. It also improves resource utilization as the smaller data packets are aggregated into larger ones before transmission, reducing overhead in packet headers. Moreover, this approach presents a scalable and an inexpensive alternative to small cell heterogeneous networks (SCHNs) and is widely accepted in the literature.

As the devices have the option of choosing between single- and two-hop transmissions, and

each provides different network performance, deciding to which AP devices connect is of critical importance and should be carefully designed such that resource utilization is improved while reducing the energy consumption of the devices. This problem is known as the AP association (APA) problem. APA is similar to the traditional radio access technology (RAT) association problem in SCHNs, where a device has the option of connecting to a small cell, a macro cell, or a WiFi AP [80]; a topic of much current interest in LTE and 5G [3]. However, the main difference between APA and RAT association is the fact that, devices in RAT association always transmit their data to the network in a single hop whereas, in APA, they have the choice of using a two-hop option. The difference introduces dependency between different APs in terms of resource utilization, since DAs are considered as cellular users in this case and share the cellular resources with directly connected devices [58].

Consequently, in this part, we aim at developing a novel APA method that jointly enhances network performance and reduces the energy consumption of the devices, while catering for the minimum throughput requirement in the network (different from part two of the work,, we do not consider delay as part of the QoS requirements). We use a centralized approach to formulate the APA problem as a mixed integer non-liner programming problem and develop a game theory based heuristic algorithm to solve it. To the best of our knowledge, besides the very recent work by Ibrahim *et al.* [58], the APA problem is new and is yet to be studied in depth. The main contributions of this part of the work are listed as follows:

- Motivated by the findings from parts one and two of this work [90,92], that data aggregation can alleviate congestion of the RACH channel in cellular IoT applications with massive number of devices, we propose to use data aggregation as a cost effective way to improve resource utilization and reduce energy consumption, while catering for the different QoS requirements. The cellular network is overlaid with DA nodes that help off-load some data traffic from the devices, leading to less congestion at the RACH channel. With proper user association, power control, and resource allocation, the DA infused network can be optimized to provide energy efficient access to a massive number of IoT devices;

- We consider joint optimization of AP association, efficient resource utilization, and device transmission power control in a DA infused cellular network. The objective is to maximize the number of QoS satisfied devices in service, while minimizing their total transmission power. We consider the limitations on the maximum allowable transmit power of the devices the heterogeneity of different APs in terms of available resources. We consider the performance dependency between the APs which is the main difference between the traditional RAT association problem in SCHNs and the APA problem considered here;

- The original multi-objective joint optimization problem is mixed-integer nonlinear and non-convex in nature, requiring discrete device-AP associations as well as discrete channel assignments, while the transmission power of the devices is continuous within a range. We propose a novel algorithm, referred to as *PAUSE* (**P**ower control, resource **A**llocation, **U**ser association, QoS **S**atisfaction and **E**nergy consumption optimization) algorithm, to effectively solve the optimization problem based on coalition formation game and difference between two concave functions programming (D.C. programming) optimization theories;

- Using the proposed algorithm, we present three different case studies that help shed light the impact of DA density and available resources on the number of supported devices and their energy consumption.

## 1.5 Thesis objectives and outline

The objective of this PhD research is to develop radio spectrum and energy efficient cellular channel access for the future anticipated massive cellular IoT applications that consist of a large number of battery powered devices producing small data packets with stringent quality requirements. In order to achieve this objective, the following steps have been taken.

1. **Random Access Channel Procedure**: In the first step, we study the RACH procedure of the traditional cellular network and model its performance using tools from stochastic geometry. We use the models to test the performance of RACH under different device densities and power control factors. Using the results of this part, we draw on some conclusions about the limitations of the RACH procedure that give insight on possible solutions;

2. **NOMA-enabled cluster-assisted Network**: Based on the first part, we notice that as the number of devices contending for the RACH preambles increase, the overall performance of the network decreases. Thus, one way of solving the problem is to decrease the number of simultaneous access requests. However, as some devices have stringent delay and throughput requirements, using time spread may not be feasible. Thus, in the second step, we propose to use data aggregation were devices are bundled into clusters each with a cluster head to relay the data. This approach has the advantage of decreasing the contention over the RACH preambles, however, without careful design, in-cluster congestion may arise, leading to long delays and increased energy consumption. Hence, in this part, we borrow tools from stochastic geometry and queuing theory to model and analyze the two-hop data aggregation enabled network where we identify the overall system delay and

throughput performance. Using the models, we draw on some conclusion on the relation between the number of clusters formed and the performance of the proposed two-hop network. We also show the advantages brought on by data aggregation compared to direct connection via the RACH procedure. Further, as a step towards future networks, we investigate the potentials of using NOMA for in-cluster transmission as a way of improving resource utilization and supporting larger number of devices. We dive into the pros and cons of NOMA-enabled clustering and draw some conclusions that give insights on the design process of optimized two-hop data aggregation networks that are cost effective, delay aware and energy efficient, suitable for the anticipate future massive cellular IoT applications;

3. **Load balancing with data aggregation**: In this part, we examine a more realistic two-hop network made up of a layer of macrocell BSs overlaid by DAs. Different from the first and second parts, devices can connect directly to BSs or to DAs that relay their data in a two-hop manner. Performance of both routes is dependent as DAs are cellular devices that share resources at the BS. In this network, device-AP association decisions become of critical importance as they determine the overall network performance. Accordingly, we tackle this issue by formulating the energy and resource efficient device association problem as an optimization problem to determine the optimal association decisions that results in the maximization of QoS satisfied supported devices.

The rest of this thesis is organized as follows: Chapter 2 reviews related research works and provides the necessary background. Chapters 3, 4, and 5 entail the preceding three steps taken towards the general objective of the Ph.D research respectively [90–92]. Finally, Chapter 6 provides conclusions and future work.

# Chapter 2

# Background and Related Works

In this work, we study massive cellular IoT communications in two-tier cellular network. We focus on two main research areas: massive access and user association. In this chapter, we first introduce the massive access problem in the cellular network and we list the available solutions. We then narrow the literature survey into three research topics: RACH uplink performance analysis, channel access in clustered IoT applications, and load balancing via user association.

## 2.1 Massive access problem in cellular networks

To gain access in cellular networks, a device uses the four step random access channel procedure known as the RACH procedure. The four steps are: 1) A device chooses (generates) a preamble from a fixed set of preambles and transmits it to its desired BS over the PRACH; 2) Once the BS receives and decodes the preamble, it sends a random access response (RAR) to the transmitting device on the physical downlink control channel (PDCCH); 3) Once the device receives its RAR, it transmits its control messages on the allocated uplink resource blocks on the physical uplink control channel (PUCCH); 4) The BS then sends an acknowledgement indicating successful association. The number of available preambles is limited; thus, as the number of contending devices grows, the probability of multiple devices choosing the same preamble increases. Many existing works assume that colliding devices on a preamble will be denied access, making the performance of the RACH similar to slotted ALOHA. Thus, under massive simultaneous access requests, the RACH performance degrades immensely. This problem is referred to by many as the RACH overload problem.

As shown in Figure 2.1, many RACH overload control methods have been proposed. They can be categorized into four categories: class barring, resource allocation, data aggregation, and others. First, in class barring, devices are allowed to contend over the PRACH with a probability, $\rho_{acb}$. The choice of the contention probability can be fixed or dynamically adjusted based on a certain metric [35,70]. Prioritized class baring can also be achieved in a similar manner via using multiple contention probabilities for different classes based on their service requirements [76,

118]. Second, in resource allocation, different applications with different delay requirements are allocated different sets of orthogonal preambles to contend over. Traditionally, the resources are split such that delay tolerant applications are allocated smaller number of preambles compared to delay intolerant applications. The preambles can be split in a preset or a dynamic manner [73]. Third, in data aggregation, devices are grouped into clusters with powerful nodes assigned as cluster heads. A cluster head acts as a data aggregator and a relay that transmits the data on behalf of its cluster members. Data aggregation has the advantage of lowering the number of devices contending over the PRACH, at the cost of multi-hop communication [48]. Data aggregation is not new as it has been the main enabling technology for WSN (see [1] and the references within). Nonetheless, IoT applications present new challenges due to the massive number of devices involved. Yet, existing works adapts many of the proposed solutions for WSNs, but with a focus on aspects such as scalable channel access and energy efficiency [87]. Last, other RACH overload control mechanisms introduced hybrid solutions that combine techniques from the other three categories [67,120], whereas, others tackled the problem from different prospective such as user association [52].

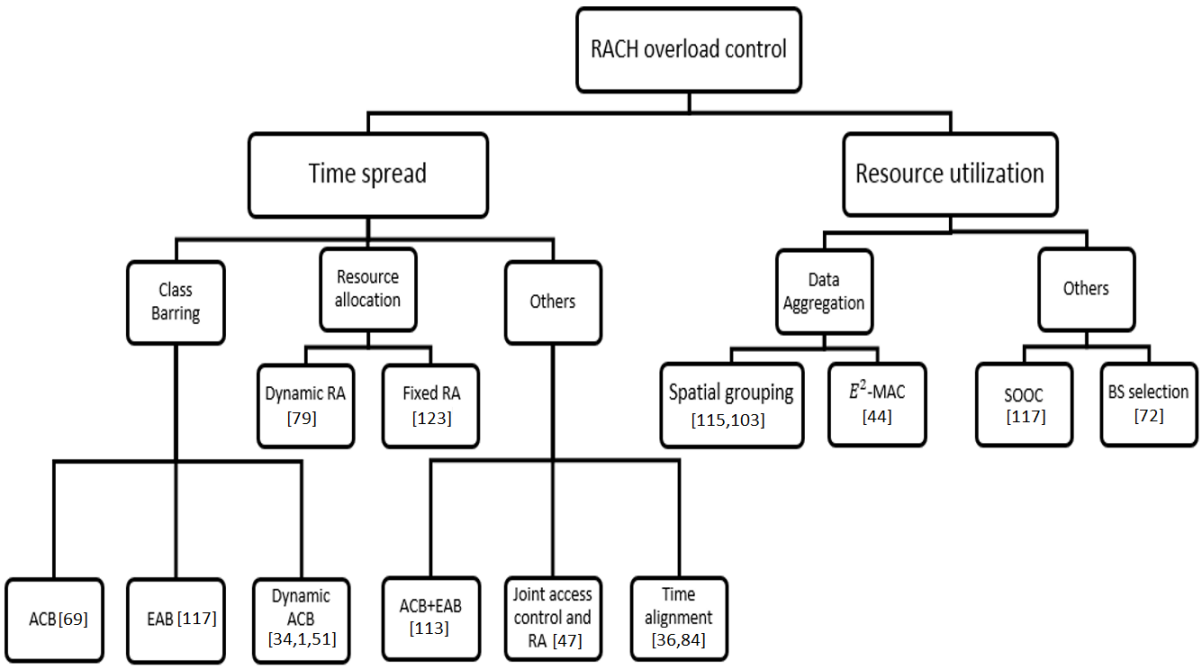


Figure 2.1: Survey of RACH overload control mechanisms

Most RACH overload control mechanisms have focused on maximizing the instantaneous RACH success probability through temporal spreading of the access requests [52]. While the results of those studies are promising, they do not tackle the core of the problem and they lead to extensive delays [70]. On the other hand, data aggregation recognizes the inherent inefficiency of the RACH procedure when supporting IoT applications. IoT devices generate

short data packets that may be smaller than the overhead introduced by the RACH procedure. Consequently, by combining individual data packets into larger ones, data aggregation not only maximizes the instantaneous RACH success probability, but also enhances the overall resource utilization [48, 87]. Thus, data aggregation is seen as a promising solution to the overload problem, but at the cost of complex multi-hop communication [84, 88, 89].

## 2.2 Uplink performance analysis in cellular networks

To understand the limitations of the RACH procedure, deriving its success probability is crucial. With the improvements in computational abilities of the cellular BSs and the introduction of various multi-user detection techniques, interfering devices can still complete the RACH procedure successfully if their transmitted preambles are received with SINR above a certain threshold [45, 106]. Thereupon, to study the success probability of the RACH procedure, the SINR should be modeled. Since RACH is a type of uplink transmission, in this section, we review the available studies conducted on modeling the SINR in uplink cellular networks.

In the uplink of cellular communication, devices usually transmit in an uncoordinated manner; hence, both inter-cell interference and intra-cell interference should be considered [37, 85], especially in the RACH procedure [46]. Interference is a function of propagation distance and channel conditions. The location of devices, being randomly deployed, complicates the SINR as the relative locations of interfering devices should be taken into account. Thanks to its ability to incorporate the spatial characteristics of networks, stochastic geometry has been used to model the statistical conditions of many complex wireless networks [51]. Many researchers have used it to analyze the coverage/outage probabilities through characterizing the SINR [9, 31, 37, 38, 113]. The accuracy of the produced models depends on two main aspects: spatial abstraction of the network topology (i.e. locations of BSs and devices) and SINR characterization [10, 11]. However, each of these aspects is network dependent and should be properly adjusted to the network under consideration for accurate results.

Most existing works focus on modeling the downlink performance due to its simplicity, where the network topology can be accurately abstracted using a simple Poisson point process (PPP) [17]. Closed form expressions for the coverage probability and rate have been derived [9,30,38,62]. However, analysis of the uplink transmission is more complex due to the uplink power control, spatial correlation of interfering devices, device hardware limitations, and heterogeneity of the devices [37, 85, 111]. Although modeling the locations of the devices in uplink transmission using the simple PPP traces unrealistic scenarios that ignores the association policy employed, many have used it to simplify the analysis [77, 112]. The use of PPP is justified in the case when

devices employ power control such that they form a Voronoi tessellation with a single BS located in the Voronoi cell of each device [95]. To produce more accurate results, other uplink network topology abstractions have been used. For instance, in [15, 111], the locations of the devices are modeled as a form of Voronoi perturbed lattice process which retains some of the inherent location dependencies among interfering devices, yet produces intractable mathematical results. Although no closed form expressions are found, the results give insights on the complexity of the uplink analysis with power control. There exists more accurate topological abstraction to model the positions of interfering devices [32, 37]. These models account for the spatial correlation between interfering devices and the serving BSs, at the cost of model complexity and intractability [18].

In most existing studies on uplink modeling in single-tier cellular networks, it is assumed that transmissions to the same BS is across a set of orthogonal channels such that no intra-cell interference is present. Consequently, the SINR, used to conduct the performance analysis, considers only inter-cell interference which greatly simplifies the derivations [37, 111]. However, when studying the performance of the RACH procedure, this assumption is not valid as multiple devices can choose the same preamble and transmit at the same time anywhere in the network, leading to both intra- and inter-cell interference which need to be incorporated when characterizing the SINR. Accordingly, since intra-cell interfering devices tend to be closer to each other than inter-cell interfering devices, they appear more clustered than dispersed. Thus, more general point processes have been used to capture this cluster-spread structure such as point cluster processes (PCP) [113]. Only a few papers have considered the impact of RACH procedure on the uplink performance of single-tier cellular networks [45, 46].

Additionally, there are limited studies on the uplink transmissions in multi-tier networks [37, 111]. Different from single-tier networks, in multi-tier networks, devices are usually located in an area with multiple BSs to associate with. This adds another dimension of randomness that should be accounted for when modeling uplink transmissions [38]. For example, in [37], devices are assumed to associate to the BSs according to their average link quality; in [111], the association police is based on maximum downlink biased power. The association policy affects the interference geometry by changing the minimum separation distance between a device and its interferes; therefore, the spatial distribution of the BSs cannot be easily abstracted [38,39,51]. Many simplifications and assumptions have to be made before choosing an appropriate point process for topological abstraction [38]. To the best of our knowledge, the RACH performance is yet to be studied in multi-tier networks under a massive access.

## 2.3 Channel access in clustered IoT communications

Data aggregation has gained interest as a possible way of supporting massive IoT applications where devices are grouped into clusters and the cluster heads relay the aggregated data on their behalf [48, 84]. This is done in a multi-hop fashion where the devices form links to their serving cluster head which then connects to the BS via the RACH procedure. Data aggregation has the advantage of reducing congestion at the BS access channel as less number of devices will contend over the PRACH. Devices can access their cluster heads in two ways: OMA and NOMA [109,119]. Table 2.1 lists the main advantages and disadvantages of OMA and NOMA. Most of the existing works on data aggregation in massive IoT communications have focused on optimizing the delay and energy efficiency of OMA based channel access solutions [48,64,84,105], and only a few have considered NOMA [7]. These studies prove the potential of NOMA as a method for supporting massive connectivity in future cellular networks in both uplink and downlink directions [108].

Table 2.1: Comparison between OMA and NOMA

| | **Advantages** [109, 119] | **Disadvantages** [109, 119] |
|---|---|---|
| **OMA** | Low complexity receiver<br>Simple MAC design | Low spectral efficiency<br>Limited number of supportable devices<br>Extensive access delays under high loads |
| **NOMA** | High spectral efficiency<br>High number of supportable devices<br>Enhance user fairness<br>Low latency<br>Possible differentiated service | High receiver complexity<br>Increased network interference<br>High sensitivity to channel uncertainties |

OMA based medium access control (MAC) protocols have been well studied and implemented in many applications. Essentially, OMA based MAC protocols can be classified into three categories: contention based, contention free, and hybrid protocols [27]. Contention based protocols, such as ALOHA, perform poorly under heavy loading conditions, making them less favored when supporting a massive number of devices. On the other hand, contention free protocols, such as time division multiple access (TDMA) and frequency division multiple access (FDMA), provide efficient spectrum utilization under heavy loads, but are inefficient under light loads. Furthermore, for the large dynamic IoT networks, a huge amount of signaling would be required to perform contention free based channel access. As for hybrid MAC protocols, they combine features from both contention based and contention free protocols. The performance of hybrid protocols highly depends on the transmission frame design [93]. As OMA allows devices to share the resources orthogonally, it likely will lead to a bottleneck when supporting massive number of devices.

On the other hand, NOMA has recently gained attention as a potential medium sharing

scheme to improve the spectral efficiency and increase the number of supported devices in the future 5G networks [108]. In general, NOMA is capable of achieving higher user capacity and providing better performance in both uplink and downlink transmission than traditional OMA [14, 71]. The key idea of NOMA is to exploit other domains, such as code or power domains, to allow multiple access, such that multiple users can share the same time/frequency/code orthogonal resources [27]. Successive interference cancellation (SIC) is used at the receiver to decode the superimposed message and separate signals from different users [104]. There are two crucial aspects required for a successful NOMA: effective interference management and low receiver complexity.

Several NOMA strategies, such as power domain NOMA and code domain NOMA are studied and their potentials and challenges in both uplink and downlink cellular networks are identified [27]. Devices, multiplexed on the same orthogonal resource block, generate an aggregate message which is decoded by the receiver in a sequential manner starting with the signal of highest power while dealing with the rest as noise [104]. The complexity of the receiver is directly proportional to the number of devices transmitting on the same resource block. The optimal performance gains of NOMA, in terms of throughput, energy efficiency and receiver complexity, are achieved when multiplexing only signals from two devices per resource block. Signals from more than two devices can be multiplexed with each other to increase device capacity, but at the cost of degraded network performance and increased receiver complexity [5, 7, 123].

To multiplex more users while not compromising performance, user pairing strategies can be used. User pairing is a way by which devices are grouped together according to specific characteristics, such as transmission power, location, and throughput requirements, such that the co-channel interference is managed and the aggregated message is correctly decoded with a high probability. Pairing near users with far ones helps improve the decoding success rate and thus achieves better rate performance and increases the number of supportable devices [4, 14]. If devices employ power control, random user pairing can be coupled with power allocation to introduce controlled interference [33]. With proper power allocation among paired users, NOMA improves the overall network performance in terms of user fairness and throughput [34, 114, 123].

User pairing and power control in the downlink is easier than in the uplink as, in the downlink, the BS can act as a centralized entity overseeing the pairing/allocation process to achieve certain network performance maximization. Although the BS can still control user pairing in the uplink, it would require extensive amount of signaling to know the locations and the channel conditions of the devices, which may limit its gains. Furthermore, devices engaging in IoT communication are massive in number and may not be able to efficiently collaborate with each other to maximize

transmission performance in a distributed manner [122]. Yet, NOMA can be paired with OMA-based access schemes and still achieve performance gains if properly designed distributed power control schemes are used [7, 24, 124].

## 2.4 User association in multi-tier cellular networks

Conventional device association rules are based on maximum received signal strength (max-RSS) or maximum signal to interference ratio (max-SIR) which are no longer the optimal association for the multi-tier architecture of future 5G multi-tier networks. Due to the disparity in the transmit power of the different tiers, using the max-RSS rule leads to a huge amount of devices associating with the macro BS as it is transmitting at the highest power, leaving smaller cells, such as pico- and femto-cells, with low or no load. This creates undesirable load imbalance in the cellular network [12, 26].

In release 10 by the 3GPP group, a device association scheme that is based on cell range expansion (CRE) through biasing is proposed [62]. In CRE, the received power at the devices from the small cell BSs are artificially altered by multiplying them with a tier specific biasing factor that shall motivate more devices to associate with them. The effectiveness of this method has been proven in recent publications [49, 82]. However, devices, forced to associate with the small biased cells, experience a strong interference in both the uplink and downlink directions [61]. Therefore, load balancing through biased device association has to be carefully designed as to optimally trade-off load balancing and network throughput by carefully selecting the bias values [23, 56].

To choose the proper biasing factors for optimal device association, most existing works rely on modeling utility based on a certain network performance, and optimizing it to achieve a certain objective. To optimize user association, game theory, combinatorial optimization and stochastic geometry are the most widely adopted tools for this purpose. Game theory is a mathematical modeling tool used to analyze the interaction of multiple players and help assign strategy such that an equilibrium (referred to as nash equilibrium) is reached. At equilibrium, a player that changes its strategy will degrade the utility of others. In the user association realm, players can be the BS [78, 79] or the users [50] or both [102] and the objective is to find the best user association strategies to optimize network performance subject to the chosen metric. Game theory is suitable for designing distributed algorithms which suits the user association problem with mobility or randomness in traffic. It also has the advantage of not introducing high overhead [117]. However, in multi-tier networks supporting massive IoT applications, game theoretic approaches may be computationally expensive and time consuming. Furthermore, the

players may not have the same objective, leading to irrationality among them [68].

Utility maximization using combinatorial optimization relies on defining an indication matrix the elements of which refer to the device-BS association. The resultant problem is NP-hard in general which may prove to be computationally prohibitive even for medium size networks. It can be made into a convex problem by using relaxation methods and then invoking techniques such as Lagrangian dual analysis [16]. However, due to the discrete nature of primal combinatorial optimization, relaxation may lead to a duality gap between the primal and dual problems [81, 107]. Furthermore, the resulting solutions usually assume centralized architecture which is not scalable and not necessarily present in IoT applications [47].

In stochastic geometry based analysis, the network is assumed to obey a certain point process, which captures the network properties [11, 38]. Based on the specific properties of the selected point process, analytic expressions can be derived for the interference, coverage probability, outage probability, and other network parameters [30, 32]. Stochastic geometry can yield tractable models irrespective of the size of the network being analyzed, making them suitable for modeling massive IoT communications [51]. Stochastic geometry is mainly used for modeling but not to find the biasing values; only a few have used it to solve the user association optimization problem [77]. This is because a closed form expression of the performance metric is not guaranteed especially in complex multi-tier networks, leading to possibly an NP-hard user association optimization problem. Out of the three methodologies, stochastic geometry emerges as a powerful tool suitable for analyzing massive IoT communications. However, it has to be combined with distributed decision algorithms to produce effective user association algorithms [107].

Table 2.2 summarizes most of the work done in the UA area based on the key performance metrics (PM) used. As shown, most of the published work have focused on downlink analysis as it was assumed that uplink and downlink are coupled; hence, optimizing the performance in one direction would optimize the performance in the other. Nevertheless, according to recent literature, decoupling uplink and downlink UA leads to large potential gains in network performance [105]; hence, techniques used to optimize downlink UA may not be optimal for uplink UA. Yet, uplink analysis are more complex due to the large number of devices and the lack of coordination.

## 2.5 Summary

In this chapter, we discuss existing studies on uplink performance analysis in cellular networks, channel access in clustered IoT communications, and load balancing through user association.

Table 2.2: User association in HetNets

| Ref. | Tool | UA mechanism | Main PM | Secondary PM | Direction | # of Tiers |
|---|---|---|---|---|---|---|
| [32] | Stochastic geometry ( modeling) | Max RSS | Coverage probability | Spectrum efficiency and slight improvement in QoS. No fairness is considered except in 115 where optimization is done to the logarithmic utility to observe fairness. | DL | K-tier |
| [22] | | Max-RSS + spectrum partitioning | | | | 2-Tier |
| [62] | | Biasing | | | | K-Tier |
| [110] | | | | | | K-tiers |
| [13] | | Biasing + spectrum partitioning | | | | K-tier |
| [103] | | | | | | K-tier |
| [77] | | | | | | K-tier |
| [37] | | Truncated channel inversion PC | Outage probability | Spectral efficiency | UL | 1- & K-tier |
| [102] | Game Theory (Distributed) | User association | Spectrum Efficiency | Fairness and some form of QoS provisioning | DL | 2-tier |
| [79] | | | | | UL | 2-tier |
| [50] | | User association | | | UL | K-tier |
| [55] | | | | | | 1-tier |
| [47] | Combinatorial Optimization | Orthogonal resource allocation (Centralized) | Spectrum Efficiency | Fairness | DL | 2-tier |
| [107] | | pricing-based biased user association (Distributed) | Energy Efficiency | QoS provisioning +fairness | DL | K-tier |
| [76] | | Biased small cells +spectrum partitioning (Centralized) | Spectrum Efficiency +Energy Efficiency | QoS Provisioning | DL | K-tier |
| [97] | | max-RSS + minimum required transmission power (Centralized) | Energy Efficiency | QoS provisioning + some fairness | UL | 2-tier |

The focus of this PhD research is to develop a general framework that enables IoT communication on the future multi-tier cellular networks. The current cellular network is deemed inefficient for IoT communications due to massive access problem, arising due to the RACH procedure. However, this claim is not well supported with appropriate analytic models. Thus, proper mathematical frameworks are needed to analyze the RACH performance of the cellular network under heavy loading conditions. Stochastic geometry is seen as a powerful modeling tool that have been extensively used for uplink modeling in large scale networks. However, most of the existing models have considered inter-cell interference only. Due to the nature of the RACH procedure, both inter- and intra-cell interference are present. Therefore, new analytic frameworks are needed as the applicability of the current models may be limited.

Data aggregation is seen as a potential solution for alleviating RACH congestion but at the cost of introducing multi-hop communications. Existing solutions have focused on minimizing energy consumption by increasing the number of the clusters or by ignoring the delay constraints of the IoT applications. Since data aggregation uses two-hop communication (i.e. device→DA→BS), optimizing both transmission phases is crucial for the overall network performance. To maximize the performance of the second phase, the number of DAs should be minimized as not to congest the PRACH. However, decreasing the number of DAs may lead to undesirable in-cluster congestion as the number of devices per cluster will increase. Given the limited uplink resources available at the DAs, to support larger number of devices while keeping

the number of clusters to minimum, new scalable channel access mechanisms are needed such that the two-hop communication performance is not compromised.

Finally, in multi-tier networks, load balancing is necessary for optimal resource utilization. However, existing load balancing mechanisms focus on data hungry applications and do not take into consideration the unique characteristics of IoT communications. Furthermore, most of the solutions have focused on downlink transmission to optimize network performance. On the other hand, IoT applications produce mostly uplink data; thus, new specifically designed load balancing schemes, suitable for the unique features of IoT applications, are urgently needed.

# Chapter 3

# RACH Performance Analysis for Cellular IoT Applications

In traditional cellular networks, devices use the RACH protocol to first associate with their desired BS before they are allowed to transmit their data in single hop fashion. The RACH protocol is a contention based channel access mechanism where devices compete over a limited number of preambles. Devices randomly choose and transmit their preambles to their desired BSs. For a preamble to be successfully received and its owner be granted access, it should be received with SINR above a certain threshold. Devices simultaneously transmitting the same preamble interfere with each other, leading to possible in-accurate decoding at the BS and resulting in denial of service. Devices that are denied service have to retry associating in next random access opportunity (RAO) which leads to an increase in the delay incurred as well as increases energy consumption. Quantifying and modeling the interference of the RACH protocol is of critical importance especially that it is expected to remain the default association mechanism in future cellular network. By providing accurate models, performance analysis can be conducted and useful conclusions can be made to guide the design of future MTC technology. Accordingly, the objective of this part of the work is to model the performance of the RACH procedure under massive number of devices. We start by identifying and modeling the in-network interference. It should be noted that we assume that the network is isolated, i.e., external interference from other communication platforms is assumed to be zero.

## 3.1 Problem description

Interference is a function of the power of the transmitted preambles by all interfering devices. As the transmitted preambles experiences path-loss attenuation and channel gain, the propagation distance of a transmitted preamble exhibits a random component. Further, since preambles are chosen randomly, this creates a random set of interfering devices on a preamble with random propagation distances among them. This adds a complexity component when modeling

the in-network interference. Capturing the randomness in the propagation distances and the transmission power among interfering devices on a preamble is crucial for accurate modeling of the SINR and for proper analysis of the performance the RACH protocol. Therefore, the spatial characteristics of the network plays an important role in determining the SINR of any received preamble in the network and shall be taken into consideration when evaluating performance of the RACH protocol. In an attempt for developing a comprehensive RACH performance model that takes into account various system variables, in this Chapter, we use stochastic geometry to analyze the association success probability of the RACH protocol in a large-scale single-tier cellular network that supports a large number of homogeneous IoT devices. Inspired by the methodology in [111], we consider FPC and biased device association.

The rest of this Chapter is structured as follows. In Section 3.2, we describe the system model under consideration. In Section 3.3, the spatial characteristics of the network are used to characterize the interference components in the network. Section 3.4 presents a step-by-step derivation of the association success probability for the RACH protocol. Section 3.5 discusses the numerical and simulation results and summarizes the main findings. This study is then concluded in Section 3.6.

## 3.2 System model

In this section, we first describe the spatial characteristics of the network under consideration, and briefly summarize the RACH association process and the collision process among contending devices. The physical channel model is then introduced with details on the power control mechanism employed by the devices in the network.

### 3.2.1 Spatial description

As shown in Figure 3.1, consider a large-scale single-tier cellular network with identical BSs that are spatially distributed in $\mathbb{R}^2$ according to a PPP $\Phi_b = \{b_0, b_1, b_2, ...\}$ with component density $\lambda_b$ where $b_j$ denotes the location of the $j^{th}$ BS. The network supports a massive number ($\mathcal{D}$) of stationary identical devices that are uniformly distributed across the plane. Devices can be in one of two states: active or inactive. An inactive device has no data to transmit, and stays in the sleep mode in which it switches off its communication module to save energy. Inactive devices are not associated with any BS. Once a device has data to transmit, its state changes from inactive to active. It then attempts to associate with its desired BS according to the maximum received signal strength (max-RSS) association rule. A device receives orthogonal preamble messages from multiple BSs. It decodes the preambles, determines the ID of each BS,

and measures the downlink received power from each. It then starts the association process with the BS from which it received the highest downlink power [59]. Devices are allowed to associate with only one BS at a time. Once a device completes its data transmission, it switches off its communication module and terminates the BS association; thus, new association is needed every time that a device has data to transmit.

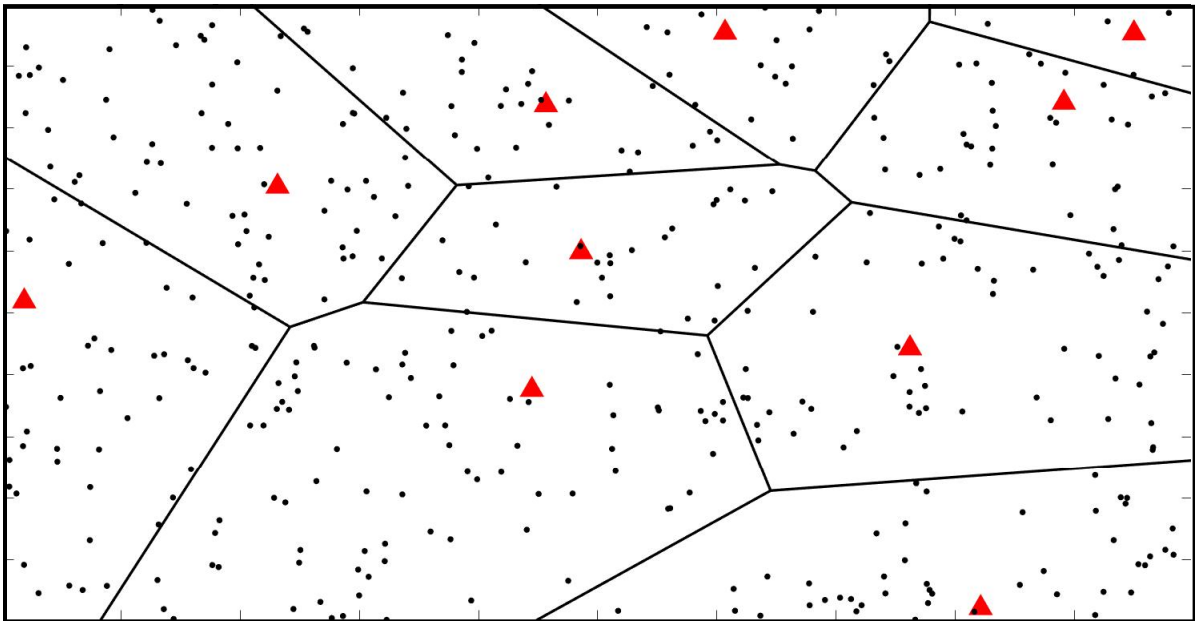

Figure 3.1: General system model for IoT communication. The dotes "." refer to the devices, the triangles "△" refer to the BSs.

Time is divided into equal transmission frames, referred to as *rounds*, each of length $T$, as shown in Figure 3.2. Each *round* is divided into $f$ sub-frames, and each sub-frame has two slots: RAO and a transmission period. A batch of devices becomes active at the beginning of each *round*. Active devices are spatially distributed according to a PPP $\Phi_d = \{d_1, d_2, ...\}$ with device density $\lambda_d$ ($>> \lambda_b$) where $d_i$ denotes the location of the $i^{th}$ active device. A *round* is sufficiently long such that there will be no backlogged data at the devices from previous *rounds*.

A device goes through two stages before it successfully completes its data transmission [28,70]. The first stage is the association stage where the RACH protocol is invoked and it takes place during the RAO slot of a sub-frame. The second stage is the transmission stage which takes place during the transmission slot of a sub-frame. In the transmission stage, successfully associated devices are allocated orthogonal resources over which they transmit their data [28]. Here we give a brief description of the first stage.

Every time a device becomes active, it uses the RACH protocol to associate with a BS [28, 70, 89]. As shown in Figure 3.3, in the first step of the RACH, a device randomly chooses (generates) a preamble and transmits it to its desired BS on the shared PRACH. In response,

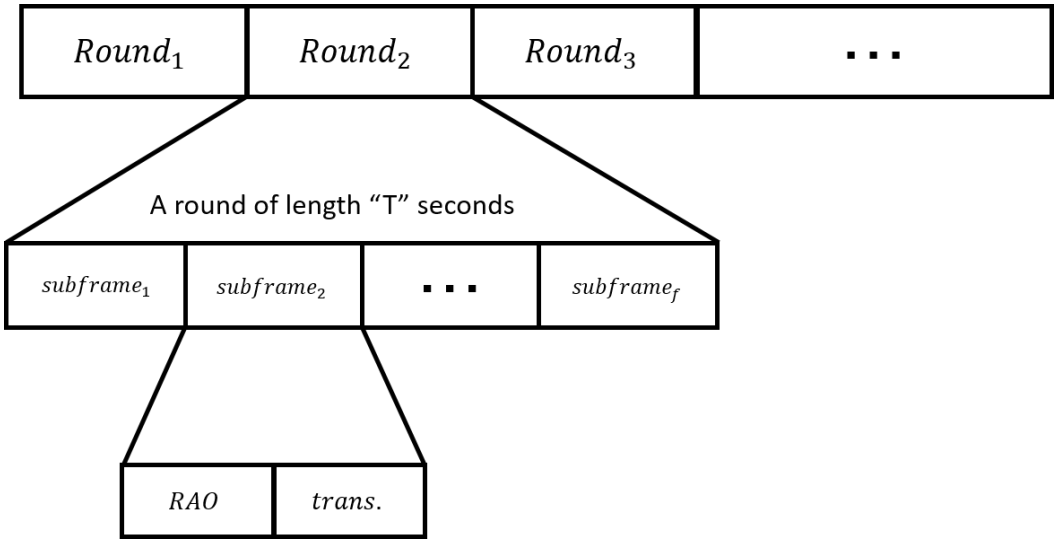


Figure 3.2: Channel access time frame structure

in the second step of the RACH protocol, the BS broadcasts a RAR message that consists of the detected preamble index corresponding to the sequence sent by the device, along with other control information. In the third step, the device uses the control information in the received RAR to synchronize with the BS and awaits its dedicated uplink resource blocks (RBs) on the physical uplink control channel. In the dedicated RBs, the device transmits a connection request using a radio resource control message which contains identity information as well as the amount of uplink resources required for it to transmit its payload. If the control message is successfully received by the BS, an acknowledgment is sent to the transmitting device indicating the assigned uplink orthogonal resources. Correct reception of the acknowledgement by the device successfully completes its association [28, 89].

All devices share a fixed number of preambles, $N$ ($<< \mathcal{D}$). All the preambles are equiprobable to be selected, each with probability $\beta = 1/N$. Preambles are transmitted in broadcast manner and thus are heard by all nearby BSs in the network. Therefore, all devices choosing the same preamble and transmitting at the same time will interfere with each other at their respective desired BSs. A device can successfully associate with its desired BS, if its preamble is received with SINR above a threshold $\tau_b$ at the BS. In Section 3.3, we characterize the SINR in the network in details. For ease of notation, in the rest of this chapter, we use $\tau$ instead of $\tau_b$ to indicate the SINR threshold for successful reception of a signal at the BS. It should be noted that the following analysis is applicable under the assumption of homogeneous devices that have the same QoS requirements and same hardware. In the case of heterogeneous devices, more component of the RACH procedure should be taken into account; however, this is out of the scope of this research.

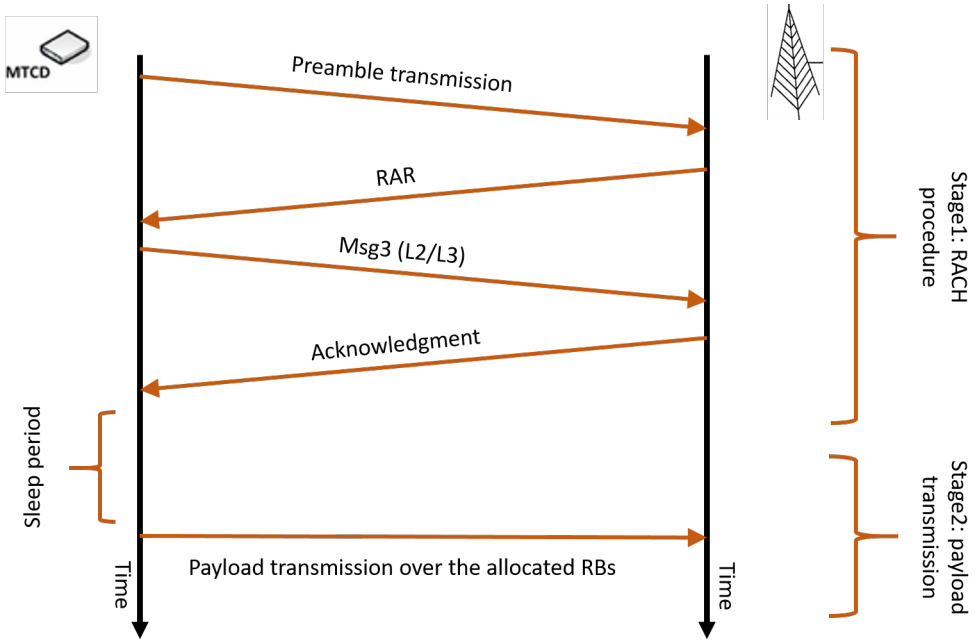


Figure 3.3: Two stage transmission process

### 3.2.2 Transmission model

Transmitted signals in both uplink and downlink experience propagation attenuation according to a general power-law path-loss model. The signal power decays at rate $D^{-\alpha}$, where $D$ is the propagation distance and $\alpha$ is the path-loss exponent. Consider a Rayleigh fading channel that introduces a random instantaneous power gain, $g$, which follows an exponential distribution with unity mean (i.e. $g \sim \exp\{1\}$). Channel gains are assumed to be distance independent as well as independent of each other and identically distributed (i.i.d.). The channel also introduces additive white Gaussian noise (AWGN) with received noise power, $\sigma^2$.

Devices employ FPC mechanism by which they adjust their transmission power to compensate partially or fully for the propagation attenuation. Accordingly, denote the device transmission power before performing FPC by the nominal power, $q_d$, which is the same for all devices [37]. For a device located at $d$ and transmitting its preamble to a BS located at $b$, the received power at the BS is $Q_d^r = q_d||d-b||^{-\alpha}g$ in the absence of noise, where $||d-b||$ denotes the Euclidean distance between the device and the BS. Now, using FPC, the device adjusts its transmission power such that it can compensate for the attenuation due to propagation loss. Therefore, the transmission power of the device after FPC is $Q_d^t = q_d||d-b||^{\epsilon\alpha}$, where $\epsilon \in [0,1]$ is the power control factor and is a design parameter that is fixed for all devices in the network. If $\epsilon = 1$, the device achieves full power control; if $\epsilon = 0$, the device does no power control. Hence, the power received at the BS is $Q_d^{FPC} = q_d||d-b||^{(\epsilon-1)\alpha}g$. In order for a device to transmit its signal, its maximum allowable transmission power should be high enough for the fractional

path-loss compensation to be achievable. For the purpose of this Chapter and to be able to develop a model and test the limits of the network under consideration, we do not consider the maximum transmission power limitation of the devices; and assumption that has been made by many in the literature for the sake of simplicity [37, 111]. However, the uplink transmission limitation is considered in Chapter 5.

## 3.3 SINR characterization

Since the set of preambles is shared by all the BSs in the network, as shown in Figure 3.4, there are two types of interference in the network: intra-cell interference ($I_{in}$) and inter-cell interference ($I_{out}$). Note that hexagonal cells are used in the figure for illustration purposes only. Due to the randomness in the channel, channel fading and path-loss attenuation, each cell will have a random coverage area that follows Voronoi cell tessellation. Intra-cell interference is due to transmissions from devices choosing the same preamble while associated with the same BS, whereas inter-cell interference is due to transmissions from devices choosing the same preamble and associated with different BSs.

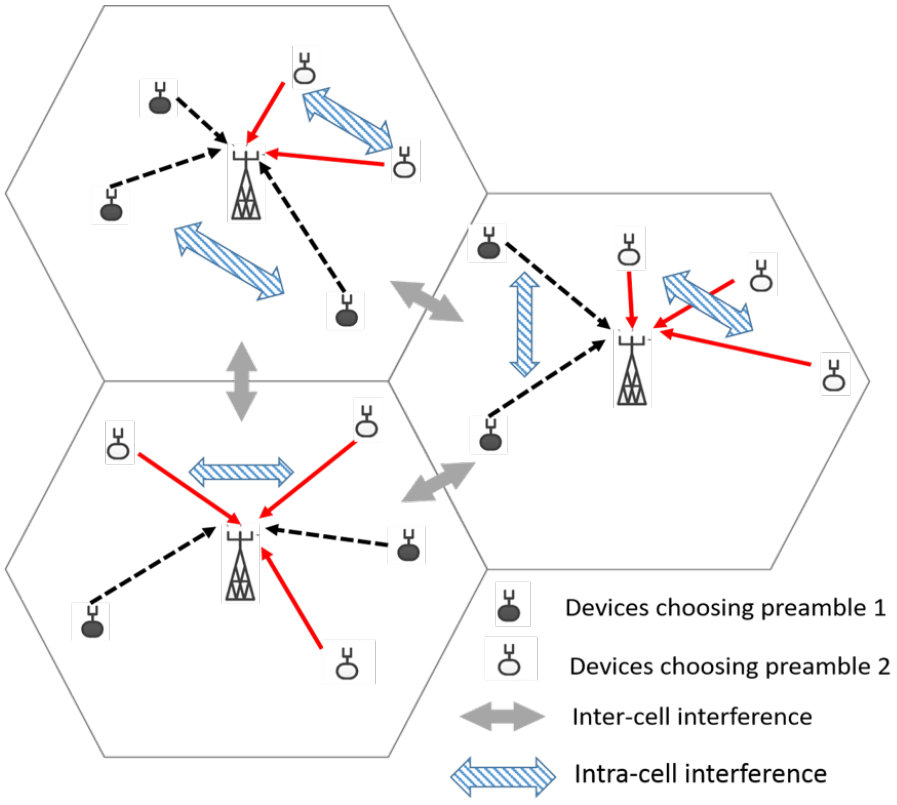


Figure 3.4: Types of interference in a single tier network

Since devices choose their preambles randomly and independently, interfering devices on the $n^{th}$ preamble can be spatially modeled by a thinned PPP $\Psi_n = \{d_1^n, d_2^n, ...\}$ with device density

$\lambda_n = \beta\lambda_d$, where $d_i^n$ denotes the location of the $i^{th}$ device interfering on the $n^{th}$ preamble for $n = 1, 2, ..., N$. Note that all active devices have to choose a preamble to transmit; therefore, $\mathbf{\Phi}_d = \cup_{n=1}^{N} \mathbf{\Psi}_n$. Also, let $\mathbf{\Psi}_{nj} = \{d_{j1}^n, d_{j2}^n, ....\}$ denote the location set of intra-cell interfering devices on the $n^{th}$ preamble and are associated with the $j^{th}$ BS for $j = 0, 1, 2, ...$ and $n = 1, 2, ..., N$. The combination of those sets constitutes the total location set of all devices choosing the $n^{th}$ preamble across the network (i.e. $\cup_j \mathbf{\Psi}_{nj} = \mathbf{\Psi}_n$)).

### 3.3.1 SINR mathematical representation

Without loss of generality, we focus on a typical BS and its associated typical device. The coordinates are shifted such that the BS is located at the origin (i.e., $b = (0, 0)$). According to Slivnyak's theorem [51], the presented analysis can be generalized for any generic device-BS pair located anywhere in the network. Both the typical BS and the typical device are indexed by zero (i.e., $j = 0$ and $i = 0$). That is, the typical device associating with the typical BS via transmitting the $n^{th}$ preamble is located at $d_{ji}^n = d_{00}^n$. For ease of notation, let $\text{BS}_0$ denote the typical BS and $\text{device}_0$ denote the typical device.

Now, let $\Xi_{ji}^n$ denote the SINR level of the $n^{th}$ preamble transmitted by the $i^{th}$ device that is associated with the $j^{th}$ BS. Hence, for $\text{device}_0$ associating with $\text{BS}_0$, the SINR of the received preamble is given by

$$\Xi_{00}^n = \frac{q_d ||d_{00}^n||^{(\epsilon-1)\alpha} g_0}{I_{in} + I_{out} + \sigma^2}. \tag{3.1}$$

In (3.1), $I_{in}$ and $I_{out}$ are given by

$$I_{in} = \sum_{d_{0i}^n \in \mathbf{\Psi}_{n0}/\{d_{00}^n\}} q_d ||d_{0i}^n||^{(\epsilon-1)\alpha} g_i \tag{3.2a}$$

$$I_{out} = \sum_{j \neq 0} \sum_{d_{ji}^n \in \mathbf{\Psi}_{nj}} q_d ||d_{ji}^n - b_j||^{\epsilon\alpha} ||d_{ji}^n||^{-\alpha} g_i \tag{3.2b}$$

where $g_i$ is the exponential channel gain with unity mean experienced by the $i^{th}$ device, $\mathbf{\Psi}_{n0}/\{d_{00}^n\}$ is the set of locations of intra-cell interfering devices associated with $\text{BS}_0$ excluding the location of $\text{device}_0$, $||d_{ji}^n||$ denotes the Euclidean distance between the $i^{th}$ inter-cell interfering device and the origin, and $||d_{ji}^n - b_j||$ is the Euclidean distance between the $i^{th}$ inter-cell interfering device and its serving BS (the $j^{th}$ BS). Note that $||d_{00}^n||$, $||d_{0i}^n||$, $||d_{ji}^n||$, and $||d_{ji}^n - b_j||$ are all random distances as they depend on the locations of the device and the BS. In the rest of this chapter, these random distances are referred to as link lengths as defined below.

### 3.3.2 Link length characterization

A link length refers to the Euclidean distance between a device and a BS. Since devices transmit in the uplink direction in a broadcast manner, the signal from each device is received by all BSs. Therefore, there is a link length between each device and each BS in the network. These link lengths can be classified into 4 types. As shown in Figure 3.5, the 4 types of link lengths are: 1) between device$_0$ and BS$_0$ ($R_{00}^n = ||d_{00}^n||$), 2) between the $i^{th}$ intra-cell interfering device and BS$_0$ ($R_{0i}^n = ||d_{0i}^n||$), 3) between the $i^{th}$ inter-cell interfering device and the $j^{th}$ BS (i.e. its serving BS) ($Y_{ji}^n = ||d_{ji}^n - b_j||$), and 4) between the $i^{th}$ inter-cell interfering device and BS$_0$ ($X_{ji}^n = ||d_{ji}^n||$).

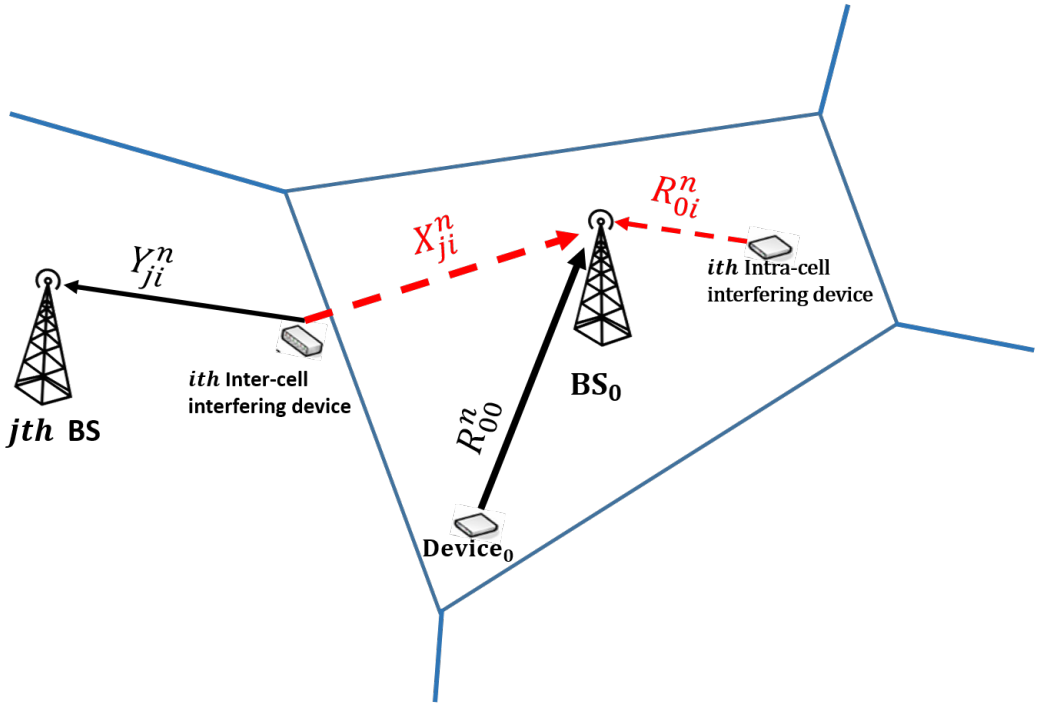


Figure 3.5: Link length of inter-cell and intra-cell interfering devices

Let $Z$ denote the distance between any device and its serving BS. Since BSs are distributed according to a PPP with density $\lambda_b$ and devices are allowed to associate with only one BS at a time, $Z$ has cumulative density function (CDF) given by the void probability of the PPP defined as [51]

$$F_Z(z) = \exp\left(-\pi\lambda_b z^2\right). \tag{3.3}$$

Accordingly, from (3.3), the probability density functions (PDF)s of $R_{00}^n$, $R_{0i}^n$ and $Y_{ji}^n$ can be shown to follow a Rayleigh distribution defined as [51, 111]

$$f_Z(z) = 2\pi\lambda_b z \exp\left\{(-\pi\lambda_b z^2)\right\}. \tag{3.4}$$

As for $X_{ji}^n$, it is possible that an inter-cell interfering device is located closer to BS$_0$ than

some of the devices located inside the Voronoi cell of $BS_0$. This phenomena is due to the Voronoi tessellation formed by the the spatial abstraction of the BS locations using PPP (see Figure 3.5). Thus, there is no exact way of differentiating between inter- and intra-cell interfering devices. One way of establishing the segregation is by assuming that a device is inter-cell interfering if at least one BS, other than $BS_0$, is located within a ball of radius $h$ centered at the device. The probability of this event is the complement of the void probability of the PPP, given by $1-\exp\left(-\pi\lambda_b h^2\right)$. Accordingly, the set of inter-cell interfering devices can be spatially modeled using non-homogeneous PPP with density $\lambda_{nI} = \lambda_n(1-\exp\left(-\pi\lambda_b h^2\right))$ with respect to the typical BS. Under this assumption, there is no exact distribution for $X^n_{ji}$ [51].

## 3.4 RACH performance analysis

For a device to successfully associate with its desired BS, its transmitted preamble should be received with an SINR above threshold, $\tau$. Let $\Theta(\Xi^n_{ji})$ denote the association success probability of the $i^{th}$ device with the $j^{th}$ BS via transmitting the $n^{th}$ preamble. Since preambles are orthogonal, the following analysis is valid for any preamble. Without loss of generality, we focus our analysis on $\text{device}_0$ associating with $BS_0$ located at the origin via transmitting the $n^{th}$ preamble. The association success probability of this typical device is defined as

$$\Theta(\Xi^n_{00}) = P(\Xi^n_{00} > \tau) = P\left(\frac{q_d||d^n_{00}||^{(\epsilon-1)\alpha} g_0}{I_{in} + I_{out} + \sigma^2} > \tau\right) \tag{3.5}$$

As $g_0$ is exponentially distributed with unity mean, (3.5) can be written as (see Appendix A.1)

$$\Theta(\Xi^n_{00}) = E_{||d^n_{00}||}\left[\exp\{-\frac{\tau}{q_d}||d^n_{00}||^{(1-\epsilon)\alpha}\sigma^2\}\mathscr{L}_{I_{in}}\{\frac{\tau}{q_d}||d^n_{00}||^{(1-\epsilon)\alpha}\}\mathscr{L}_{I_{out}}\{\frac{\tau}{q_d}||d^n_{00}||^{(1-\epsilon)\alpha}\}\right] \tag{3.6}$$

where $\mathscr{L}_{I_{in}}\{.\}$ and $\mathscr{L}_{I_{out}}\{.\}$ denote the Laplace transforms of intra-cell and inter-cell interference components respectively.

For notational simplicity, let $s = \frac{\tau}{q_d}||d^n_{00}||^{(1-\epsilon)\alpha}$ such that, given $||d^n_{00}||$, (3.6) can be rewritten as the conditional success probability given by

$$\Theta(\Xi^n_{00}|d^n_{00}) = \exp\{-s\sigma^2\}\mathscr{L}_{I_{in}}\{s\}\mathscr{L}_{I_{out}}\{s\}. \tag{3.7}$$

To calculate the Laplace transforms in (3.7) and find the association success probability, the number of devices in one Voronoi cell needs to be calculated first. Since the number and locations of active devices at the beginning of each transmission *round* are random, the number of devices in each cell is also random. In the following, we derive an approximation for the

average number of devices within each Voronoi cell using device density and area of the coverage cell.

### 3.4.1 Average number of devices in a cell

Let $M_0$ denote the total number of active devices located in the Voronoi cell of $BS_0$. Denote $P(M_0 = m)$ as the probability mass function (PMF) of $M_0$. Since the locations of active devices follow a PPP with density $\lambda_d$ and the BSs form a Voronoi tessellation, the number of active devices inside the coverage area of any BS follows a conditional Poisson distribution given by [51, 125]

$$P(M_0 = m|V) = \frac{(\lambda_d V)^m e^{-\lambda_d V}}{m!} \tag{3.8}$$

where $V$ is the area of the Voronoi cell of the BS which is random in nature. The exact distribution of $V$ is not known [51]; however, it can be approximated by a generalized Gamma distribution given by [41]

$$f_V(v; \lambda_b, c) = \frac{\lambda_b^c v^{c-1} e^{-\lambda_b v}}{\Gamma(c)} \tag{3.9}$$

where $c = 3.575$ is a constant defined for the Voronoi tessellation in $\mathbb{R}^2$ [96] and $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t}\, \mathrm{d}t$. Therefore, the unconditional PMF of $M_0$ is approximately given by

$$\begin{aligned} P(M_0 = m) &= \int_0^\infty \frac{(\lambda_d v)^m e^{-\lambda_d v}}{m!} f_V(v; \lambda_b, c)\, \mathrm{d}v \\ &\approx \frac{\lambda_b^c \lambda_d^m \Gamma(m + c)}{\Gamma(c)(\lambda_b + \lambda_d)^{m+c} m!}. \end{aligned} \tag{3.10}$$

Accordingly, the average number of active devices in a Voronoi cell in this network is given by (see Appendix A.2)

$$\bar{M}_0 = \frac{\gamma \zeta \Gamma(c + 1)}{(1 - \zeta)^{(c+1)}} \tag{3.11}$$

where $\gamma = \frac{\lambda_b^c}{\Gamma(c)(\lambda_b+\lambda_d)^c}$ and $\zeta = \frac{\lambda_d}{\lambda_b+\lambda_d}$.

Due to the uniformity of the device distribution, the average number of active devices in the Voronoi cell of any BS is the same, i.e., $\bar{M}_j = \bar{M}_0$ for $j = 1, 2, ....$ For simplicity, we assume that the number of devices in each cell is high enough such that there is at least one device on each preamble. As devices randomly choose the preambles and each preamble is equiprobable to be chosen, we assume that active devices will be divided equally across the $N$ available preambles.

To insure that the number of active devices is integer, define $\bar{M}_j^*$ as the rounded-up average number of active devices in a cell to the nearest multiple of $N$ (i.e., $\bar{M}_j^* = \lceil \frac{\bar{M}_0}{N} \rceil$ for $j = 0, 1, 2, ...$). Accordingly, let $\bar{M}_j^n$ denote the average number of devices interfering on the $n^{th}$ preamble and

associated with the $j^{th}$ BS for $j = 0, 1, 2, ...$ and $n = 1, 2, ..., N$ . Then, $\bar{M}_j^n$ is given by

$$\bar{M}_j^n = \frac{\bar{M}_j^*}{N} \quad \forall n \; and \; \forall j. \tag{3.12}$$

Since the average number of devices in each cell is the same and the preambles are equiprobable to be chosen, for simplicity of notation, the superscript and subscript are omitted and thus for the rest of the analysis, we have $\bar{M} = \bar{M}_j^n$ to denote the average number of intra-cell interfering devices on any preamble and associated with any BS. Using the above results, expressions for the Laplace transforms of the interference components in (3.7) are derived next.

### 3.4.2 Interference Laplace transforms

As a further simplification, we assume that each BS will have exactly $\bar{M}^* = N \cdot \bar{M}$ active devices located within its Voronoi cell at the beginning of each *round*. Although this assumption is dubious, it will help reduce the complexity of the Laplace transforms. The accuracy of this assumption is validated by means of simulations as will be shown in the results section.

Intra-cell interference is due to devices transmitting the same preamble while attempting to associate with the same BS. For $\text{device}_0$ associated with $\text{BS}_0$, intra-cell interference is a function of the distance from devices located in the Voronoi cell of $\text{BS}_0$ to the origin. The Laplace transform of $I_{in}$ is approximately given by (See Appendix A.3)

$$\begin{aligned}
\mathscr{L}_{I_{in}}\{s\} &= E_{I_{in}}[\exp\{-sI_{in}\}] \\
&= E_{g_i, s, ||d_{0i}^n||}\left[\exp\{-sq_d \sum_{d_{0i}^n \in \Psi_{n0}/\{d_{00}^n\}} ||d_{0i}^n||^{(\epsilon-1)\alpha} g_i\}\right] \\
&\approx \left(\int_0^\infty \frac{f_{R^n}(r)}{1 + sq_d r^{(\epsilon-1)\alpha}} \mathrm{d}r\right)^{\bar{M}}
\end{aligned} \tag{3.13}$$

where $f_{R^n}(r)$ is the i.i.d. PDF of $R_{0i}^n$ (for different $i$ values) given by (3.4).

Inter-cell interference is due to devices located outside the Voronoi cell of $\text{BS}_0$. Inter-cell interfering devices adjust their power to compensate for the path-loss attenuation based on the distances to their serving BSs. For simplicity, assume that the locations of inter-cell interfering devices are modeled by a homogeneous PPP of density $\lambda_n$ [51]. This assumption is valid when the density of the BSs is high enough, such that the area of the Voronoi cells of each BS becomes very small. With omni-directional antennas, transmissions from inter-cell interfering devices generate interference at $\text{BS}_0$. The Laplace transform of the aggregate inter-cell interference is approximately given by (see Appendix A.4)

$$\mathscr{L}_{I_{out}}\{s\} = E_{I_{out}}[\exp\{-sI_{out}\}]$$

$$
\begin{aligned}
&= E_{g_i,s,||d_{ji}^n - b_j||,||d_{ji}^n||}\left[\exp\{-sq_d \sum_{j\neq 0}\sum_{d_{ji}^n\in\Psi_{nj}} ||d_{ji}^n||^{-\alpha}||d_{ji}^n - b_j||^{\epsilon\alpha} g_i\}\right] \\
&= E_{g_i,s,Y_{ji}^n,X_{ji}^n}\left[\exp\{-sq_d \sum_{j\neq 0}\sum_{d_{ji}^n\in\Psi_{nj}} (Y_{ji}^n)^{\epsilon\alpha}(X_{ji}^n)^{-\alpha} g_i\}\right] \\
&\approx \exp\left(-2\pi\lambda_n \int_y^\infty \left[1 - \left(\int_0^\infty \frac{f_{Y^n}(y)}{1+sq_d y^{\epsilon\alpha}x^{-\alpha}}\mathrm{d}y\right)^{\bar{M}}\right] x\,\mathrm{d}x\right)
\end{aligned}
\tag{3.14}
$$

where $f_{Y^n}(y)$ is the i.i.d. PDF of $Y_{ji}^n$ (for different $i$ and $j$ values) given by (3.4).

Substituting (3.13) and (3.14) into (3.7), the final form of the general conditional success probability, as a function of $\sigma$, $\tau$, $\lambda_n$, $\alpha$ and $\epsilon$, conditioned on $s$, is given by

$$
\begin{aligned}
\Theta(\Xi_{00}^n)_{(\sigma,\tau,\lambda_n,\alpha,\epsilon|s)} &= \exp\{-s\sigma^2\}\mathscr{L}_{I_{in}}\{s\}\mathscr{L}_{I_{out}}\{s\} \\
&\approx \left[e^{-\left(s\sigma^2 + 2\pi\lambda_n \int_y^\infty \left[1-\left(\int_0^\infty \frac{f_{Y^n}(y)}{1+sq_d y^{\epsilon\alpha}x^{-\alpha}}dy\right)^{\bar{M}}\right]xdx\right)}\left(\int_0^\infty \frac{f_{R^n}(r)}{1+sq_d r^{(\epsilon-1)\alpha}}\mathrm{d}r\right)^{\bar{M}}\right].
\end{aligned}
\tag{3.15}
$$

Since the preceding analysis is applicable to any device-BS pair, let the general association success probability for any device-BS pair be denoted by $\Theta_{as}$. In (3.15), $s = \frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha}$ is a function of the random distance $||d_{00}^n||$ between the $\mathrm{BS}_0$ and $\mathrm{device}_0$. For notational simplicity, let $W$ be the random distance between $\mathrm{device}_0$ and $\mathrm{BS}_0$ such that $W$ has a PDF $F_W(w)$ given by (3.4) (i.e. $W = R_{00}^n$). Assuming an interference limited network (i.e. $\sigma = 0$), which is a valid assumption when the number of contending devices is large, the unconditional association success probability as a function of $\sigma$, $\tau$, $\lambda_n$, $\alpha$ and $\epsilon$ is given by

$$
\begin{aligned}
\Theta_{as} &= E_W\left[\Theta(\Xi_{00}^n)_{(\sigma,\tau,\lambda_n,\alpha,\epsilon|w)}\right] \\
&= \int_0^\infty \Theta(\Xi_{00}^n)(\sigma,\tau,\lambda_n,\alpha,\epsilon,w) f_W(w)\,\mathrm{d}w \\
&= \int_0^\infty \Theta(\Xi_{00}^n)(\sigma,\tau,\lambda_n,\alpha,\epsilon,w) 2\pi\lambda_b w e^{-\pi\lambda_b w^2}\mathrm{d}w \\
&\approx 2\pi\lambda_b \int_0^\infty w\left(\int_0^\infty \frac{f_{R^n}(r)}{1+\tau w^{(1-\epsilon)\alpha} r^{(\epsilon-1)\alpha}}\mathrm{d}r\right)^{\bar{M}} e^{-\left(\pi\lambda_b w^2 + 2\pi\lambda_n \int_y^\infty \left[1-\left(\int_0^\infty \frac{f_{Y^n}(y)}{1+\tau w^{(1-\epsilon)\alpha} y^{\epsilon\alpha}x^{-\alpha}}dy\right)^{\bar{M}}\right]x\mathrm{d}x\right)}\mathrm{d}w.
\end{aligned}
\tag{3.16}
$$

## 3.5 Numerical results and discussion

In this section, we present numerical results for the association success probability based on (3.16) and simulations. The association success probability is evaluated versus three major system parameters: density of interfering devices on the $n^{th}$ preamble ($\lambda_n$), the power compensation

level ($\epsilon$), and the SINR threshold ($\tau$). With the Rayleigh distribution in (3.4) for $Y$ and $R$, there is no closed form expression for (3.16); thus, we resort to numerical calculations that are validated via independent system level simulations. Table 3.1 lists the parameters used to obtain the results.

Table 3.1: System parameters used to obtain the numerical and simulation results

| **Parameter** | **Value** | **Description** |
|---|---|---|
| $N$ | 64 | Number of orthogonal preambles shared by all BSs |
| $Q_b^t$ | 43 $dB$ | Downlink transmission power of any BS |
| $q_d$ | 1 $dB$ | Nominal uplink transmission power of any device |
| $\alpha$ | 4 | Path-loss attenuation factor |
| $\lambda_b$ | 1 BS/$km^2$ | Density of BSs |
| $\rho$ | $-65$ $dB$ | Receiver sensitivity of any device |
| $\sigma$ | 0 | Noise variance |

Simulations are performed based on the system model in Section 3.2 using MATLAB for a 200 $km$ x 200 $km$ 2D plane. Each result is an average of 1000 Monte Carlo simulation runs. In the network, locations of BSs and devices are randomly generated based on their deployment densities. We use 3GPP path-loss model [25] with Rayleigh fading of unit mean. Maximum received signal strength association rule is employed such that a device is associated with the BS from which it receives the highest downlink power. Device's receiver sensitivity is set to $\rho = -65$ $dB$ such that downlink preamble messages received with power less than $\rho$ will not be detected. Devices associated with each BS are then randomly divided into $N$ groups to simulate random preamble generation. Noise variance, $\sigma$, is set to zero to enforce network isolation and mimic the interference limited behaviour when dealing with massive number of interfering devices.

Figure 3.6 shows the impact of FPC parameter $\epsilon$ on the success association probability of the RACH protocol. The results are plotted against SINR threshold $\tau$. The FPC parameter is varied between 0 and 0.8 in steps of 0.4. Device density $\lambda_d$ ($= \lambda_n \cdot N$) is set to be 2560 devices/$km^2$. As shown in the figure, for all values of $\epsilon$, as threshold $\tau$ increases, the success association probability decreases. This behaviour is expected, as fewer devices will be able to achieve high SINR levels due to the extensive interference caused by the high contention. Furthermore, for a low value of $\tau$ ($< -14$ $dB$), increasing $\epsilon$ always results in a higher success association probability. For a larger $\tau$ value, the association success probability is almost identical irrespective of $\epsilon$; in fact, a higher $\epsilon$ value results in degradation of the success association probability. Note that, the shown results in Figure 3.6 were obtained at a device density that is respectively high. Accordingly,

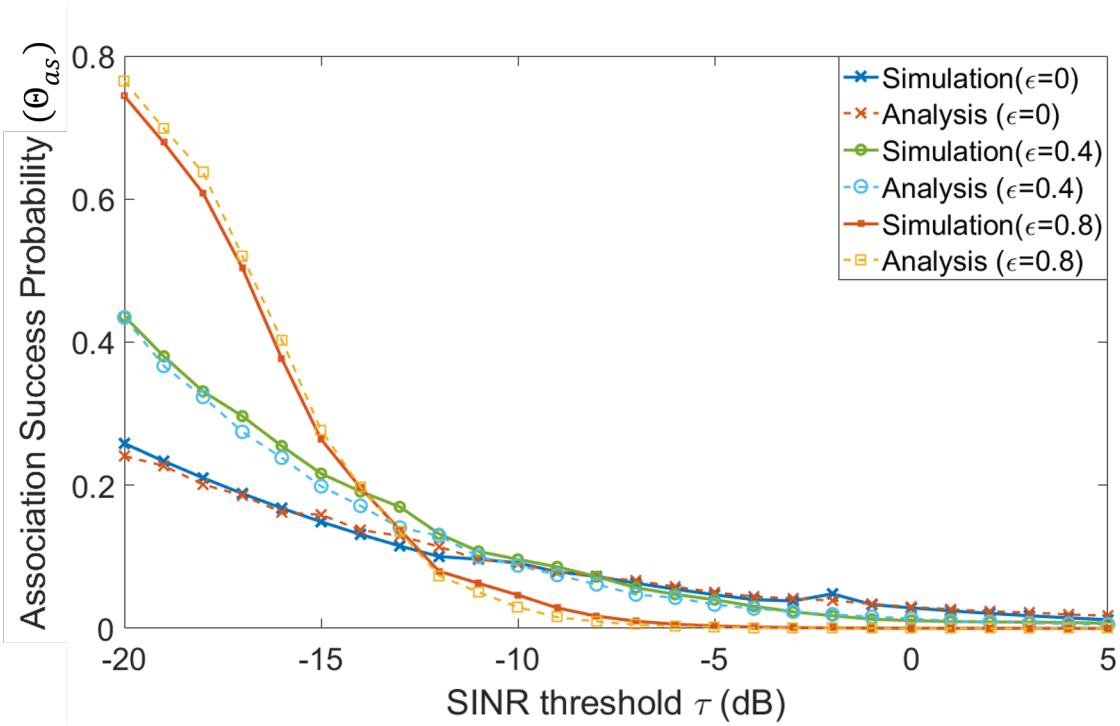


Figure 3.6: Uplink association success probability of a typical device transmitting the $n^{th}$ preamble to its serving typical BS

the results suggest that at high device densities, FPC is beneficial as long as the SINR threshold is low.

Figure 3.7 shows the relationship between increasing device density and the FPC parameter. The results are obtained by setting $\tau = -5$ $dB$, and varying device density $\lambda_n$ between 6 and 24 device/preamble/$km^2$ in steps of 2. Three $\epsilon$ values are studied: 0, 0.2, and 0.4. As shown, for the $\lambda_n$ values, increasing $\epsilon$ results in a higher success probability. However, as $\lambda_n$ increases, the improvement in success probability associated with FPC decreases as evidenced by the shrinking gap between the curves. Consequently, there is a trade-off between the density of the devices and the gain achieved by employing FPC.

In Figures 3.6 and 3.7, to be consistent with the literature, the threshold, $\tau$, varies between $-20$ $dB$ and 5 $dB$ [45]. While these SINR threshold values may seem somewhat low, we are interested in the trend rather than the actual values. In this work, we study the RACH performance under extreme loading conditions. In fact, $\lambda_n$ is set to a minimum of 6 indicating that there is at least 6 devices per BS transmitting the $n^{th}$ preamble at the same time. Under such loading conditions, the association success probability at a high SINR value will be extremely low such that the trend will not be visible. In Figure 3.8, the loading conditions are relaxed by setting $\lambda_n$ to 2 devices/preamble/$km^2$. The threshold, $\tau$, varies between $-20$ $dB$ and 5 $dB$. As expected, as the device density decreases, the success association probability increases, yet the trend with

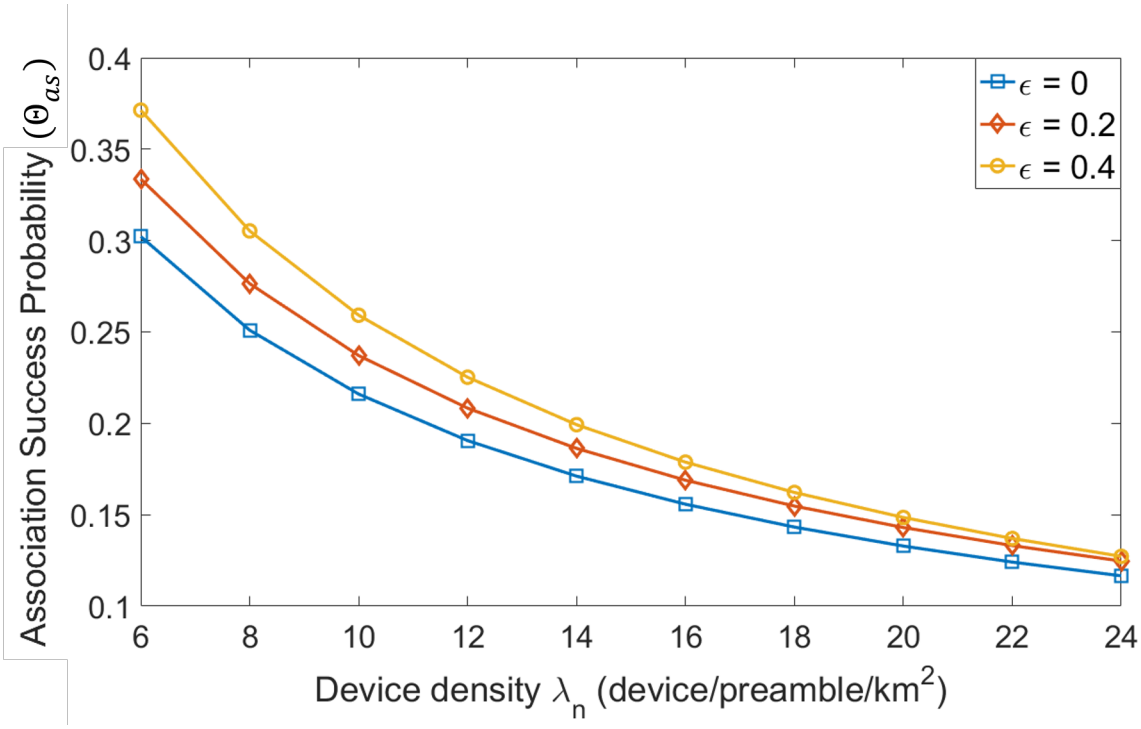


Figure 3.7: Effect of increasing device density

respect to $\tau$ stays the same. Also, the improvement due to FPC is observed.

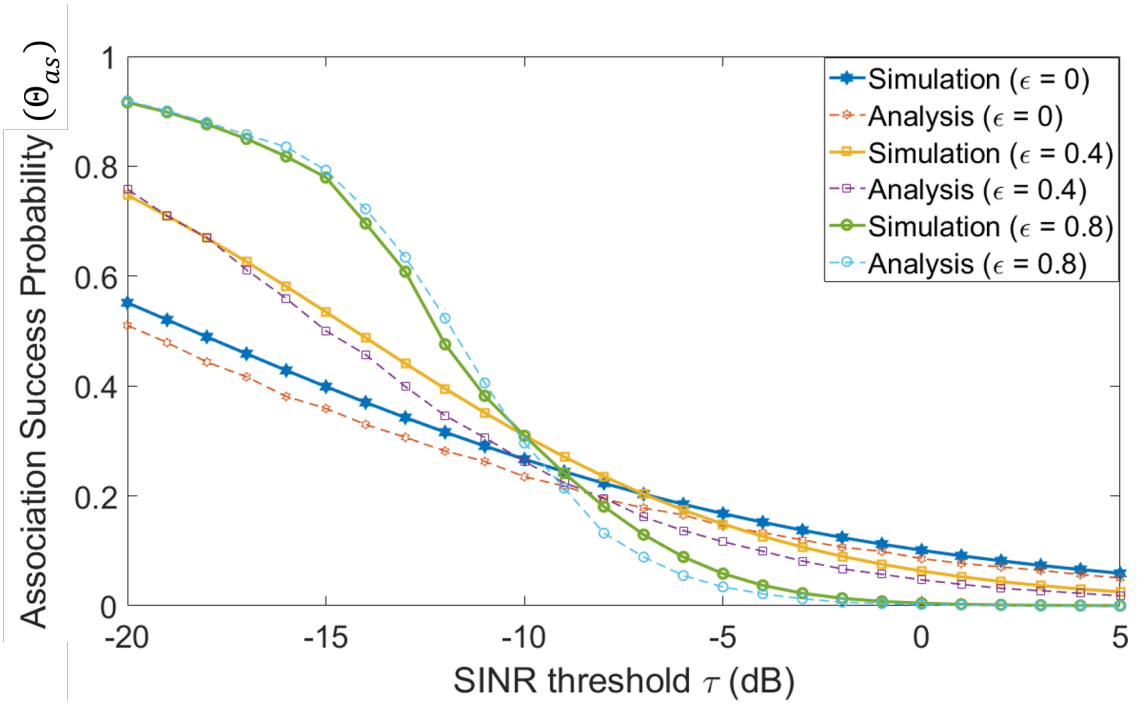


Figure 3.8: Association success probability when $\lambda_n = 2$

It is important to note that the number of competing devices changes with time within each *round*. Only a portion of the active devices will be able to successfully associate in each sub-frame, while the rest will move on to compete in the next sub-frames. This behaviour can be

characterized using the association success probability. Let $\Theta_{as}^{f}$ denote the instantaneous (e.g., in the $f^{th}$ sub-frame) association success probability, and $\lambda_d^f$ denote the progressive density of active devices in the $f^{th}$ sub-frame, which follows the relation

$$\lambda_d^{f+1} = (1 - \Theta_{as}^{f})\lambda_d^{f}. \tag{3.17}$$

Figure 3.9 shows the association success probability as time progresses (depicted by the number of RAO slots spanned within each *round*). The initial device density per preamble (i.e., $\lambda_n$) is set to 30 device/preamble/$km^2$ (i.e., $\lambda_d$ = 1920 active device/$km^2$). The SINR threshold is set at $\tau = -10$ $dB$ and no power control is used (i.e., $\epsilon = 0$). For this parameter setting, devices require 30 RAOs for all of them to successfully associate. This translates into 30 LTE sub-frames, each of length 20ms. This is under the assumptions that devices do not share the LTE-defined 64 preambles with any other applications and that the reception SINR threshold $\tau$ is somewhat unrealistically low [40]. In reality however, besides having to target higher and more realistic SINR thresholds, IoT applications will share the available preambles with other more domination human-to-human applications [28]. In fact, many studies advocate dedicating only a small portion of the preambles to IoT applications as a way of avoiding performance degradation of other applications. Having access to a smaller number of preambles leads to even a lower association success probability which translates to a higher access delay. Thus, we can conclude that the RACH protocol is indeed a bottle neck when it comes to supporting a massive number of devices requesting simultaneous access and can potentially hinder the applicability of cellular IoT communication paradigm.

Figure 3.10 shows the impact of FPC on access delay, i.e., the association success probability as time progresses. The device density is set at $\lambda_n$ = 30 device/preamble/$km^2$. Two values of the threshold $\tau$ are studied: $-15$ $dB$ and $-5$ $dB$. The FPC parameter $\epsilon$ is varied between 0 and 0.8 in steps of 0.4. As observed, at a low SINR threshold value, FPC can reduce cellular access delay; while at a high threshold value, besides increasing the energy consumption of the devices, it increases access delay. Consequently, FPC may not be an adequate solution for improving cellular access for the anticipated cellular IoT applications with a massive number of devices, as it results in higher energy wastage, increases access delay and requires operating at an unrealistically low SINR threshold value.

In this section, we study the association behavior in cellular networks under a massive number of devices. The derived model for the association success probability of the RACH protocol is analyzed and its accuracy is corroborated via system level simulations. Through extensive simulations, the derived model is shown to be accurate as evident by the close match between

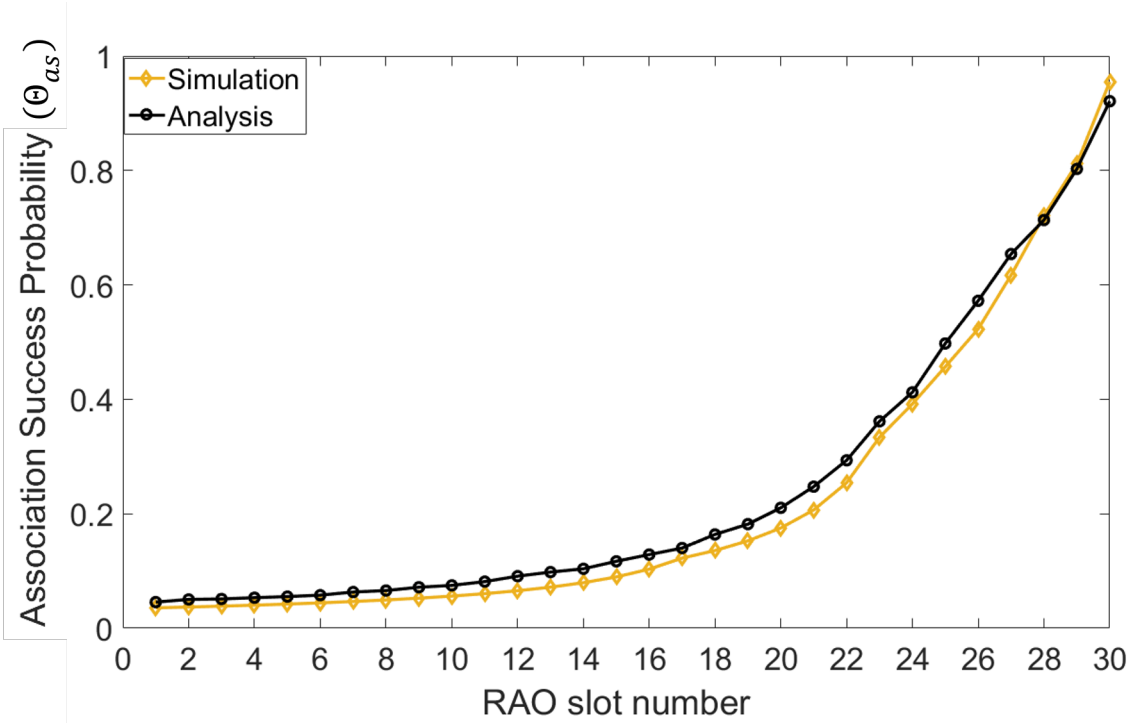


Figure 3.9: Association success probability as a function of RAO slot number ($\epsilon = 0$, $\tau = -10$, $\lambda_n = 30$)

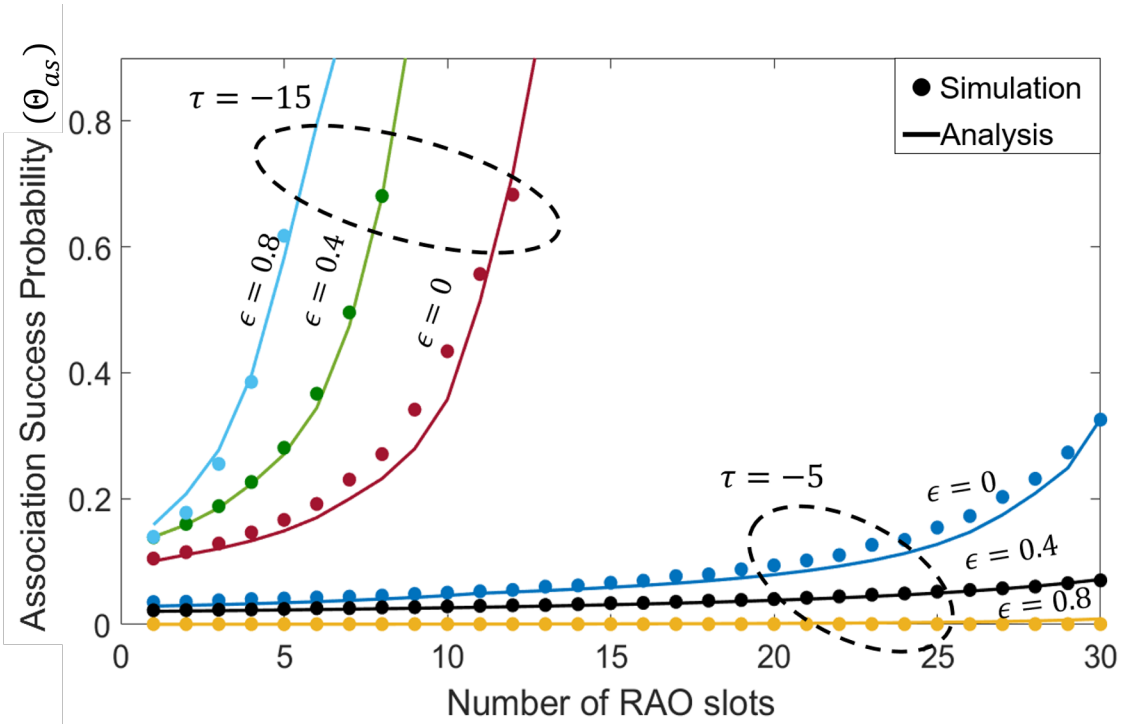


Figure 3.10: Association success probability as a function of RAO slot number, $\epsilon$ and $\tau$ ($\lambda_n = 30$)

the analysis and simulation results. Compared to the models presented in [45, 59], the derived model is more comprehensive and provides the flexibility of analyzing the RACH performance under various situations. For instance, our model can be used to analyze the performance under

different path-loss attenuation compensation levels by varying FPC parameter $\epsilon$. Furthermore, other parameters such as device density, number of available preambles, BS density and SINR threshold can be varied to study different network scenarios. Nonetheless, inspired by the work presented in [46, 59], our model can be further generalized into a spatio-temporal model by considering data arrival dynamics at the devices, which is left as a part of our future work.

## 3.6 Summary

In this chapter, we use stochastic geometry to model the RACH instantaneous association success probability in a single-tier cellular network by modeling the SINR. The objective is to study the access performance of the traditional cellular network when supporting applications with a massive number of devices such as the anticipated massive cellular IoT communications. For a device to successfully associate with its desired BS, its RACH association preamble should be received with SINR above a certain threshold. To find the association success probability, we first abstract the network topology by spatially modeling the locations of the BSs and the active devices using two independent homogeneous PPPs, and then characterize intra-cell interference and inter-cell interference components of the preamble's SINR in the network via deriving expressions for their Laplace transforms. The final model of the SINR is used to find the expression for the association success probability. The model for the association success probability is a function of device density, number of available RACH preambles, BS density, and the FPC parameter. The analytic model is corroborated via Monte Carlo simulations conducted using MATLAB. Various network scenarios are tested by using different combinations of the variables in the model. Numerical results demonstrate that there is an advantage of using a power control mechanism as a means for enhancing RACH association probability, particularly when operating at a low SINR threshold value; however, the performance gain decreases as the threshold increases. It is also noticed that, as the number of contending devices decreases, the performance gain due to the employed power control mechanism increases. Considering these observations, we conclude that the limited number of preambles in LTE degrades the access performance as the number of contending devices increases, leading to extensive access delays that may be unacceptable for delay-intolerant IoT applications. Based on the results presented in this study, it is suggested that the current association mechanism of the cellular network may not be adequate for supporting large-scale IoT applications with a massive number of devices, which require ubiquitous energy efficient and delay aware connectivity. Thus, innovative random access should be developed accordingly to efficiently support future IoT applications.

# Chapter 4

# Two-hop NOMA-Enabled Massive Cellular IoT Communications

As identified in Chapter 3, the bottle neck when it comes to supporting a large number of devices using the traditional cellular network is the contention based channel access mechanism known as RACH Procedure. One way of reducing contention is to reduce the number of contending devices which can be done using time spread (i.e., limiting the number of devices attempting association at each RAO). While this approach is plausible, it introduces sensible delays that might revoke the delay and energy efficiency requirements of many future massive cellular IoT applications. Another way that has recently gained interest is based on data aggregation, where devices are clustered and a cluster head is used to relay the data of its cluster members in a two-hop fashion. This helps reduce the number of contending devices on the available RACH preambles. At the same time, data aggregation helps reduce energy consumption as devices transmit towards DAs that are at a relatively closer proximity to them in comparison to BSs.

While data aggregation seems to be a plausible solution for efficiently supporting future massive cellular IoT applications, it introduces another dimension of complexity by transforming the single-hop network to two-hop network. Hence, although with data aggregation, contention on the PRACH declines, the network needs to be properly designed such that the two-hop architecture is optimized to meet the delay and energy efficiency requirements of future cellular IoT applications. Besides optimal two-hop network design, efficient resource utilization mechanism that can help increase the number of supported devices over the same amount of resource shall be explored. Accordingly, in this Chapter, we attempt to examine the advantages of employing NOMA-enabled data aggregation as a potential solution for providing efficient access for a massive number of energy constrained and delay sensitive cellular massive IoT applications.

## 4.1 System model

### 4.1.1 Spatial system model

Consider a layer of cellular BSs, whose locations can be modelled by a homogeneous Poisson point process (HPPP) $\Phi_b = b_0, b_1, ...$ with density $\lambda_b$, where $b_i$ denotes the location of the $i^{th}$ BS in the network. The cellular network has a massive number of identical and battery powered IoT devices, whose locations can be modelled by an independent HPPP $\Phi_d = d_0, d_1, ...$ with density $\lambda_d$, where $d_i$ denotes the location of the $i^{th}$ IoT device in the network. Data packet generation at each IoT device follows a Poisson process with parameter $\gamma$ and each generated data packet is of size $\mathcal{W}$ bits.

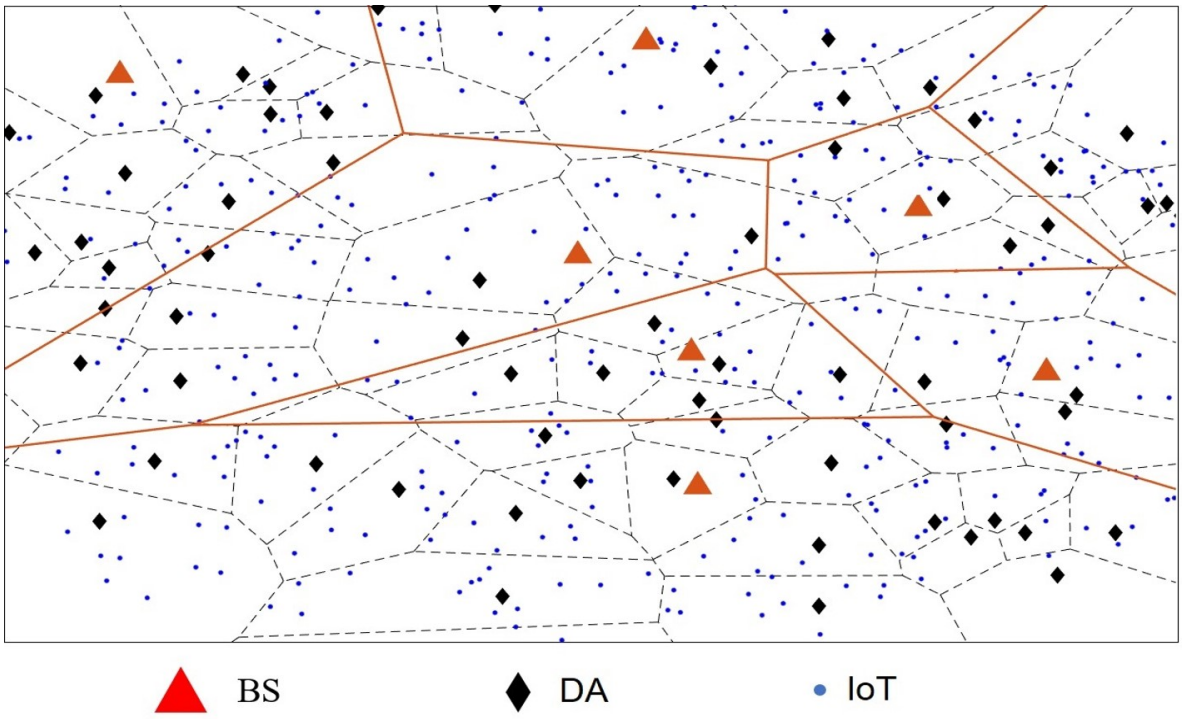


Figure 4.1: An illustration of BS, DA, and IoT device locations as well as the Voronoi tessellation formed by the coverage of the BSs (solid lines) and that formed by the coverage of the DAs (dashed lines).

The cellular network is overlaid with a layer of stationary DAs that aggregate and relay data from IoT devices to BSs. The DAs are randomly scattered across the network coverage area such that their locations can be modelled as an independent HPPP $\Phi_a = a_0, a_1, ...$ with density $\lambda_a$, where $\lambda_b < \lambda_a << \lambda_d$ and $a_i$ denotes the location of the $i^{th}$ DA in the network. Active IoT devices transmit their data packets simultaneously towards the core network only via two-hop communication, i.e., IoT devices connect to DAs, and DAs relay aggregated data packets to BSs. Maximum received signal strength association policy is used by the devices and DAs to connect

to their preferred access point. Figure 4.1 shows an example of node locations and coverage areas.

### 4.1.2 Transmission frame structure and transmission process

Time is partitioned into transmission frames of equal length $T_f$. Each transmission frame is further divided into two phases: an aggregation phase of length $T_a$ and a relay phase of length $T_r (= T_f - T_a)$. FDMA is used to split the channel between the devices and DAs into $\mathcal{N}$ sub-channels of equal bandwidth $\omega_n$. These sub-channels are used by all the DAs in the network. During the aggregation phase, devices transmit data to their serving DAs via power domain NOMA (PD-NOMA), where at most two devices from the same cluster are allowed to non-orthogonally share the same in-cluster sub-channel. The maximum number of devices that a DA can support in a single transmission frame is $M_{max} = 2\mathcal{N}$. Full channel state information (CSI) is assumed at the DAs, such that appropriate device pairing on the sub-channels is achieved in order to obtain benchmark results [7]. At the end of the aggregation phase, scheduled devices return to the pool of active devices awaiting service in the new transmission frame.

DAs employ SIC to decode the superimposed messages and cancel their mutual interference in a sequential manner. Multiplexed transmissions from devices on a sub-channel are ranked in a descending order based on their received signal strength at the DA. The DA decodes the signal from the higher ranked (stronger signal) device first, hereafter indexed by $j = h$. For the message from the higher ranked device to be successfully decoded, it should be received with an SIR above threshold $\tau_{ah}$. For the transmission from the lower ranked device, hereafter indexed by $j = l$, to be successfully decoded, two conditions should be met: i) the message from the higher ranked multiplexed device was successfully decoded, and ii) the message from the lower ranked device is received with an SIR above threshold $\tau_{al}$.

By the end of the data aggregation phase, all IoT devices switch off their communication modules and go to sleep to reduce energy consumption. The relay phase then begins where DAs transmit the aggregated data to the BS in a single hop fashion. We make a simplifying assumption that the number of DAs in any cell is equal to $\mathcal{K}$, which is the average number of DAs per cell given certain DA and BS node densities. A shared uplink channel (SUCH) of bandwidth $\Omega_b$ is dedicated for uplink transmissions from the DAs, which is divided in frequency into $\mathcal{K}$ orthogonal sub-channels of equal bandwidth. The $\mathcal{K}$ uplink sub-channels are shared by all BSs in the network. Only a single DA from the same Voronoi cell is allowed to occupy a sub-channel and thus DAs only experience inter-cell interference. A DA is considered connected to its serving BS if its transmitted data packets are received with SIR above threshold $\tau_b$.

### 4.1.3 Wireless transmission model

Transmitted signals in both phases experience propagation attenuation according to a general power-law path-loss model. The signal power decays at rate $D^{-\alpha}$, where $D$ is the propagation distance and $\alpha$ is the path-loss exponent. Consider a Rayleigh fading channel that introduces a random instantaneous power gain, $g$, which follows an exponential distribution with unity mean (i.e. $g \sim \exp\{1\}$). Channel gains are distance independent, independent of each other and identically distributed (i.e., i.i.d.). The network is interference limited due to the massive number of devices. As in our previous work [90], consider that DAs employ fractional power control (FPC) when transmitting their payload over their scheduled uplink sub-channel. Accordingly, the power received at the BS located at the origin from an associated DA located at $a$ after FPC is $Q_a^r = q_a||a||^{(\epsilon-1)\alpha}g$, where $q_a$ is the nominal transmission power of any DA in the network, $||a||$ is the Euclidean distance between the DA and the BS, and $\epsilon \in \{0, 1\}$ is the power control factor. On the other hand, IoT devices employ full power inversion such that the received power at a DA located at the origin from an associated device located at $d$ is $Q_d^r = q_d g$, where $q_d$ is the nominal transmission power of any IoT device in the network. As in previous studies, we do not consider the maximum transmission power limitation for both DAs and devices [90].

In the following few sections of this chapter, we attempt to derive mathematical models for coverage probability, end-to-end delay, and energy consumption for the preceding described system model. First, in Section-4.2, by borrowing tools from stochastic geometry and using the spatially modeling of the locations of the BSs, DAs and active devices as three independent homogeneous PPPs, we characterize interference components experienced by both the devices and the DAs. The Laplace transforms for the interference components are then derived and used to find expressions for the average device-DA and DA-BS transmission success probabilities. These probabilities are then used to derive both the device-DA coverage probability and the total coverage probability in this two-hop network. In this work, coverage probability is defined as the probability that a device is able to successfully transmit its packet to the core network via a two-hop network architecture. A device that is able to transmit its data packets successfully is referred to as a covered device. Second, using these coverage probabilities and by modeling this two-hop network as a two-stage tandem queue, in Section-4.3, we derive the arrival and departure processes from which the end-to-end average packet delay can be derived. Last, in Section-4.4, the coverage probabilities and the end-to-end delay results for the two-stage tandem queue are used to derive expressions for the average energy consumption of a typical device and a typical delay to transmit a single data packet.

## 4.2 Coverage probability

Coverage probability refers to the probability that a generated data packet from an arbitrary IoT device is successfully received by its serving BS [48]. Different from existing studies [7], the main challenge in our work is the PPP locations of IoT devices and DAs. The coverage probability consists of three parts: i) NOMA sub-channel scheduling probability - the probability of an arbitrary active device being scheduled in the current transmission frame, such that the device is able to transmit its data packet to the serving DA; ii) Device-DA coverage probability - the probability that a transmitted packet from a scheduled device is received at its serving DA with SIR above threshold $\tau_a \in \{\tau_{ah}, \tau_{al}\}$, depending on the ranking of the device; and iii) DA-BS coverage probability - the probability that a transmitted aggregated data packet from a DA is received at its serving BS with SIR above threshold $\tau_b$. The aggregation and relaying phases are correlated as the data transmitted by a DA depends on the number of scheduled devices and their SIR performance. Nonetheless, for the sake of tractability, we assume that the two phases are independent [48]. In what follows, we focus on a cluster that has its DA located at the origin (with index $i = 0$); however, the analysis is applicable to any cluster in the network. We consider the case when the network is full, i.e., at the beginning of the considered transmission frame, all devices in the network have at least one data packet to transmit. Note that, the coverage probability resembles the association success probability from Chapter 3 and hence can be used to compare single and two-hop network performance.

### 4.2.1 NOMA sub-channel scheduling probability

As mentioned in the system model, the number of devices in a cluster is random and dependent on both the density of devices and DAs. On the other hand, the number of available NOMA sub-channels is fixed and so is the maximum number of devices, $M_{max}$, that can be supported by a DA per transmission frame. Consequently, in the case that the number of devices in a cluster is greater than $M_{max}$, some devices may not be scheduled for transmission in the current frame due to insufficient resources. Thus, we define NOMA sub-channel scheduling probability as the probability that a device with data to transmit is scheduled for transmission on a sub-channel in the current frame. To determine this scheduling probability, we first characterize the distribution of the device number in a cluster. Let random variable (RV) $\mathcal{M}_a$ denote the number of devices associated with the DA over area $\mathcal{V}_a$ of the Voronoi cell. As detailed in Chapter 3 [90], by approximating the PDF of $\mathcal{V}_a$ by a generalized Gamma distribution, we obtain the probability mass function (PMF) of $\mathcal{M}_a$, given by

$$P(\mathcal{M}_a = m) = \int_0^\infty \frac{(\lambda_d v)^m e^{-\lambda_d v}}{m!} f_{\mathcal{V}_1}(v, \lambda_a, c)\ \mathrm{d}v$$
$$\approx \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a + \lambda_d)^{m+c} m!} \tag{4.1}$$

where $f_{\mathcal{V}_1}(v, \lambda_a, c) = \frac{(\lambda_a^c v^{c-1} e^{-\lambda_a v})}{\Gamma(c)}$, $c = 3.575$ is a constant defined for the Voronoi tessellation in $\mathbb{R}^2$ [96] and $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t}\ \mathrm{d}t$.

Let $\Lambda_s = \frac{M_{max}}{\max\{\mathcal{M}_a, M_{max}\}}\big|_{\mathcal{M}_a = m}$ be the conditional scheduling probability given the number of active devices in the current transmission frame, $\mathcal{M}_a = m$. Averaging over the PDF given in (4.1), the NOMA sub-channel scheduling probability is given by

$$\bar{\Lambda}_s = \sum_0^{M_{max}} \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a + \lambda_d)^{m+c} m!} + \sum_{M_{max}+1}^{\infty} \frac{M_{max}}{m} \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a + \lambda_d)^{m+c} m!}. \tag{4.2}$$

### 4.2.2 Device-DA coverage probability

To determine device-DA coverage probability, the SIR should be carefully characterized. We focus on a device from a cluster centered at the origin. As per the system model, each DA schedules its associated devices across the sub-channels in a sequential manner until a maximum of two devices are scheduled per sub-channel. The number of devices in the cluster is random, leading to randomness in the number of scheduled devices on a sub-channel. Let RV $S^n$ denote the number of devices scheduled on the $n^{th}$ sub-channel, for $n = 1, 2, ..., \mathcal{N}$. The PMF of $S^n$, $P(S^n = s)$, is given by (see appendix B.1)

$$P(S^n = s) = \begin{cases} \sum_{m=0}^{\mathcal{N}-1} (1 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 0 \\ \sum_{m=\mathcal{N}}^{M_{max}-1} (2 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m) + \sum_{m=0}^{\mathcal{N}-1} (\frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 1 \\ \sum_{m=M_{max}}^{\infty} P(\mathcal{M}_a = m) + \sum_{m=\mathcal{N}}^{M_{max}-1} (\frac{m}{\mathcal{N}} - 1) P(\mathcal{M}_a = m), & s = 2 \\ 0, & s > 2. \end{cases} \tag{4.3}$$

A device can experience two types of interference components: intra-cluster interference ($I_1$), and inter-cluster interference ($I_2$). $I_1$ can only be experience by a higher ranked device when sharing a sub-channel with a lower ranked device. On the other hand, $I_2$ can be experienced by any device regardless of its ranking. $I_2$ is composed of primary component ($I_{2h}$) due to higher ranked devices from adjacent clusters, and secondary component ($I_{2l}$) due to lower ranked devices from adjacent clusters. Locations of all higher and lower ranked inter-cluster interfering devices can be modeled by two independent and thinned PPPs $\Psi_n^h = d_{0h}^n, d_{1h}^n, ...$ and $\Psi_n^l = d_{0l}^n, d_{1l}^n, ...$, with device densities $\lambda_d^h = (P(S^n = 1) + P(S^n = 2))\lambda_a$ and $\lambda_d^l = P(S^n = 2)\lambda_a$ respectively, where

$d_{ij}^n$ denotes the location of the $j$ ranked interfering device from the $i^{th}$ cluster on the $n^{th}$ sub-channel, for $j \in \{h, l\}$, $i = 1, 2, ...$ and $n = 1, 2, ..., \mathcal{N}$. Note that, $\lambda_d^h$ follows from the fact that a device is classified as higher ranked if it is scheduled alone on a sub-channel or co-occupies the sub-channel with a lower ranked device, whereas $\lambda_d^l$ happens only when a sub-channel has more than one scheduled device. Thus, for an arbitrary device scheduled on the $n^{th}$ sub-channel and associated with the DA located at the origin, $I_{2h}$ and $I_{2l}$ are given by

$$I_{2h} = \sum_{d_{ih}^n \in \Psi_n^h} \mathbb{1}(||d_{ih}^n - a_k|| < ||d_{ih}^n||)||d_{ih}^n - a_k||^{\alpha}||d_{ih}^n||^{-\alpha} g_{ih}^{n0} \tag{4.4}$$

$$I_{2l} = \sum_{d_{il}^n \in \Psi_n^l} \mathbb{1}(||d_{il}^n - a_k|| < ||d_{il}^n||)||d_{il}^n - a_k||^{\alpha}||d_{il}^n||^{-\alpha} g_{il}^{n0} \tag{4.5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, the condition $||d_{ij}^n - a_k|| < ||d_{ij}^n||$ is to ensure that an interfering device located at $d_{ij}^n$ is closer to its serving DA located at $a_k$ than to the origin, and $g_{ij}^{n0}$ denotes the i.i.d. exponentially distributed channel gain for the link between the $j$ ranked interfering device on the $n^{th}$ sub-channel from the $i^{th}$ cluster and the origin.

For the following, we focus on a device associated with the DA located at the origin (DA at the origin is indexed by $i = 0$); however, the analysis can be generalized to any device-DA pair. Let $\Xi_{0j}^{ns}$ denote the received SIR at the origin from a $j$ ranked device located at $x_{0j}^n$ and scheduled on the $n^{th}$ sub-channel that has $s$ multiplexed devices, given by

$$\Xi_{0j}^{ns} = \begin{cases} \frac{g_{0h}^{n0}}{I_2}, & s = 1 \ \& \ j = h \\ \frac{g_{0h}^{n0}}{I_1 + I_2}, & s = 2 \ \& \ j = h \\ \frac{g_{0l}^{n0}}{I_2}, & s = 2 \ \& \ j = l \end{cases} \tag{4.6}$$

where $I_1 = g_{0l}^{n0}$ and $I_2 = I_{2h} + I_{2l}$, and $g_{0j}^{n0}$ denotes the i.i.d exponentially distributed channel gain for the link between the $j$ ranked device from the cluster indexed $i = 0$ and the origin. As we are considering full power inversion, the received power from the device of interest is $g_{0j}^{n0}$ as we assume the nominal transmission power of all the devices is set to 1 (i.e., $q_d = 1$).

As discussed, for a device to successfully transmit to the serving DA, the received signal should have SIR above $\tau_a \in \{\tau_{ah}, \tau_{al}\}$, depending on the device's rank. Accordingly, the conditional transmission success probability of a device associated with the DA located at the origin, $\Theta(\Xi_{0j}^{ns})$, given the NOMA ranking $j$ of the device and the number of scheduled devices $s$ on the $n^{th}$ sub-channel, is given by

$$\Theta(\Xi_{0j}^{ns}) = \begin{cases} P\left(\Xi_{0h}^{n1} > \tau_{ah}\right), & s = 1 \;\&\; j = h \\ P\left(\Xi_{0h}^{n2} > \tau_{ah}\right), & s = 2 \;\&\; j = h \\ P\left(\Xi_{0l}^{n2} > \tau_{al} \cap \Xi_{0h}^{n2} > \tau_{ah}\right), & s = 2 \;\&\; j = l. \end{cases} \tag{4.7}$$

Due to complexity of the interference components, there is no closed form expression of $\Theta(\Xi_{0j}^{ns})$. The probability terms in (4.7) contain the Laplace transforms for the interference components experienced by the devices; for $\alpha \neq 4$, the integrals within these Laplace transforms cannot be solved and hence a closed form expression is unattainable. On the other hand, in the case of $\alpha = 4$, a closed form expression for the conditional transmission success probability of a device can be found, given by (see Appendix B.2)

$$\Theta(\Xi_{0h}^{n1}) = \prod_{j\in\{h,l\}} \exp\left\{-\frac{\sqrt{\tau_{ah}}\lambda_d^j}{\lambda_d}\left(\frac{\pi}{2} - \arctan\left((\sqrt{\tau_{ah}})\right)\right)\right\} \tag{4.8}$$

$$\Theta(\Xi_{0h}^{n2}) = \mathcal{J}\left(\prod_{j\in\{h,l\}} \exp\left\{-\frac{\sqrt{\tau_{ah}}\lambda_d^j}{\lambda_d}\left(\frac{\pi}{2} - \arctan\left((\sqrt{\tau_{ah}})\right)\right)\right\}\right) \tag{4.9}$$

$$\Theta(\Xi_{0l}^{n2}) = \mathcal{J}\left(\prod_{j\in\{h,l\}} \exp\left\{-\frac{\sqrt{A}\lambda_d^j}{\lambda_d}\left(\frac{\pi}{2} - \arctan\left((\sqrt{A})\right)\right)\right\}\right) \tag{4.10}$$

where $A = \tau_{ah}+\tau_{al}(1+\tau_{ah})$, and $\mathcal{J} = \frac{1}{1+\tau_{ah}}$ is a result of the fact that in the case when two devices co-occupy a channel, they experience both intra- and inter-cluster interference components. It should be noted that for other $\alpha$ values, numerical evaluations can be used to compute the probabilities.

To find the device-DA transmission success probability, we need to average $\Theta(\Xi_{0j}^{ns})$ with respect to the number of scheduled devices on a sub-channel. That is, the long term device-DA transmission success probability, $\bar{\Theta}$, is given by

$$\bar{\Theta} = P(S^n = 1)\Theta(\Xi_{0h}^{n1}) + \frac{P(S^n = 2)}{2}(\Theta(\Xi_{0h}^{n2}) + \Theta(\Xi_{0l}^{n2})). \tag{4.11}$$

Accordingly, the device-DA coverage probability, $C_d$, defined as the probability that an active device is able to successfully transmit its data to the serving DA in the current transmission frame, is given by

$$C_d = \bar{\Lambda}_s \cdot \bar{\Theta}. \tag{4.12}$$

### 4.2.3 DA-BS coverage probability

A dedicated uplink channel, SUCH, is utilized for DA-BS transmissions, consisting of $\mathcal{K}$ sub-channels of equal bandwidth. Let RV $K$ denote the number of DAs in the coverage area $\mathcal{V}_b$ of a BS, which has the PMF given by [90]

$$P(K=k) \approx \frac{(\lambda_b)^c(\lambda_a)^m\Gamma(m+c)}{\Gamma(c)(\lambda_b+\lambda_a)^{m+c}m!}. \tag{4.13}$$

Thus, we have

$$\mathcal{K} = E_k[K] \approx \left\lceil \frac{\kappa\zeta\Gamma(c+1)}{(1-\zeta)^{c+1}} \right\rceil \tag{4.14}$$

where $\kappa = [(\lambda_b)^c/(\Gamma(c)(\lambda_b+\lambda_a)^c)]$ and $\zeta = [\lambda_a/(\lambda_b+\lambda_a)]$.

DAs experience only inter-cell interference. Accordingly, locations of interfering DAs on the $k^{th}$ sub-channel can be modelled by an independent PPP $\Psi_a^k = a_1^k, a_2^k, ...$ with density $\lambda_a^k = \lambda_b$. Similar to device-DA transmission success probability, for a DA to have successful transmission, the received signal at its serving BS should have SIR above $\tau_b$. Let $\Xi_a^k$ denote the received SIR at the BS located at the origin from its associated aggregator located at $a_0^k$ and transmitting on the $k^{th}$ sub-channel, given by

$$\Xi_a^k = \frac{||a_0^k||^{(\epsilon-1)\alpha} g_{00}^a}{I_a} \tag{4.15}$$

where $I_a = \sum_{a_i^k \in \Psi_a^k} ||a_i^k - b_i||^{\epsilon\alpha} ||a_i^k||^{-\alpha} g_{i0}^a$ is the inter-cell interference experienced by the DA of interest, and $g_{i0}^a$ is the channel gain for the link between the DA associated with the $i^{th}$ BS and the origin, for $i = 0, 1, 2, ...$ where $i = 0$ refers to the BS at the origin.

Let $\Theta(\Xi_a^k)$ denote the DA-BS coverage probability for the DA located at $a_0^k$ and associated with the BS located at the origin and transmitting on the $k^{th}$ sub-channel, which is given by (for details, see Appendix D in [90])

$$\Theta(\Xi_a^k) = P\left\{\Xi_a^k > \tau_b\right\} = 2\pi\lambda_b \int_0^\infty r e^{-(\pi\lambda_b r^2)} \mathscr{L}_{I_a}\left\{\tau_b r^{\alpha(1-\epsilon)}\right\} dr \tag{4.16}$$

where $\mathscr{L}_{I_a}\{x\}$ is the Laplace transform of $x$ with respect to $I_a$, given by

$$\begin{aligned}\mathscr{L}_{I_a}\left\{\tau_b r^{\alpha(1-\epsilon)}\right\} &= E_{I_a}\left[e^{\tau_b r^{\alpha(1-\epsilon)} I_a}\right] \\ &= \exp\left\{-2\pi\lambda_a^k \int_y^\infty \left(1 - \int_0^\infty \left(\frac{2\pi\lambda_a y e^{-\pi\lambda_a y^2}}{1+\tau_b r^{\alpha(1-\epsilon)} y^{\alpha\epsilon} x^{-\alpha}}\right) dy\right) x\, dx\right\}\end{aligned} \tag{4.17}$$

where $r = ||a_0^k||$ is used to denote the distance between the DA of interest and the origin, $y = ||a_i^k - b_i||$ is used to denote the distance between an arbitrary DA located at $a_i^k$ and its serving BS located at $b_i$, and $x = ||a_i^k||$ is used to denote the distance between an arbitrary DA

located at $a_i^k$ and the origin. Note that, $r$, $y$ and $x$ are RVs following Rayleigh distributions (given in (3.4)) with parameters $\lambda_b$, $\lambda_a$ and $\lambda_a^k$ respectively [90, 95].

As a result, the long term device-BS total coverage probability for a device in a NOMA-enabled two-hop network with maximum of two paired devices on a sub-channel is given by

$$C_{tot} = C_d \cdot \Theta(\Xi_a^k). \tag{4.18}$$

## 4.3 Delay performance

Average transmission delay is the expected time duration from the instant that a packet is generated at an IoT device to the instant that it is successfully received at the serving BS. In our system, the end-to-end delay is composed of two parts: the first hop delay which is the expected time that a packet spends in the queue of an IoT device until it is successfully received at the serving DA, and the second hop delay which is the expected time that the packet spends in the queue of a DA until it is successfully transmitted towards the serving BS. We model the two-hop transmission paradigm as a Tandem queue, as shown in Figure 4.2 where $\gamma$ is the packet arrival rate at a device, $\gamma_a$ is the packet arrival rate at a DA, and $\mu_a$ is the departure rate from a DA.

Packets are independently generated at each IoT device according to a Poisson process with parameter $\gamma$. Assuming infinite queue space, the arrival and departure processes at an IoT device can be modelled as M/G/1 queue, where service rate is dependent on the device's ranking on the sub-channel as well as the SIR value. On the other hand, for a DA, the packet arrival flow consists of the aggregation of all the service processes from its scheduled devices in the current transmission frame, while its service rate depends on the SIR value of its transmissions at the serving BS. Consequently, the arrival and departure processes at a DA can be modeled as G/G/1 queue under infinite queue space. In this section, we establish mathematical models for the arrival and departure processes at the devices and DAs based on queuing theory and stochastic geometry. The models are then used to find the overall average transmission delay based on the two-hop tandem queuing model.

### 4.3.1 Device-DA queue analysis

For a packet to be transmitted from a device to the DA, the device first needs to access the DA. Accessing the DA in a transmission frame means that the device is scheduled on a NOMA sub-channel and is in the DA coverage. Let $\Gamma_d$ denote the number of transmission frames needed for a device to successfully access the serving DA and transmit its data, which follows a

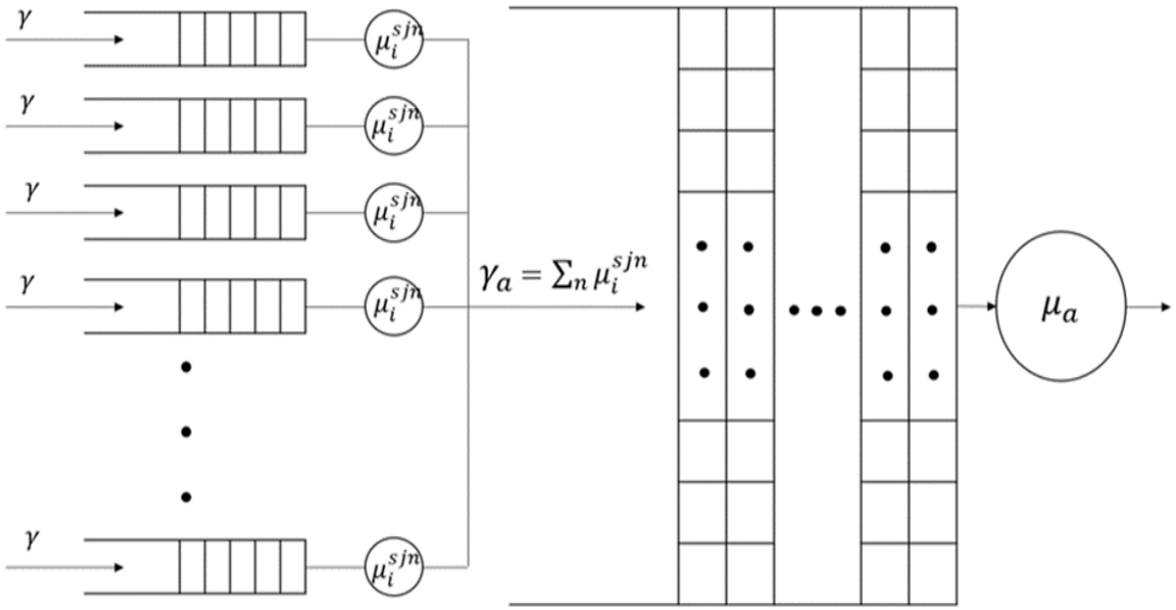


Figure 4.2: Tandem queue model of the proposed two-hop NOMA-enabled transmission network

geometric distribution with the probability of success given by device-DA coverage probability in (4.12). Thus, the average number of transmission frames needed is given by $\bar{\Gamma}_d = \frac{1}{C_d}$. As an approximation, we consider the rounded up representation of $\bar{\Gamma}_d$ and thus, packet generation and transmission at a device, over the period of $\Gamma_d^* = T_f \cdot \lceil \bar{\Gamma}_d \rceil$, can be modeled as an M/G/1 queue with packet arrival rate $\gamma$. As for the departure rate, a device, scheduled on a NOMA sub-channel, transmits its data during the aggregation period of the transmission frame during which it accessed its DA. Let $\mu_d^{sjn}$ denote the packet service rate of the $j$ ranked device scheduled on the $n^{th}$ sub-channel that has $s$ multiplexed devices during the period of $\Gamma_d^*$, given by

$$\mu_d^{sjn} = \left\lfloor \frac{T_a \mathcal{R}_d^{sjn}}{\mathcal{D}} \right\rfloor, \quad s \in \{1,2\}, \quad j \in \{h,l\}, \quad n = 1,2,...,\mathcal{N} \tag{4.19}$$

where we use the flooring function to avoid fractions and to emphasize that a device transmits a packet if and only if the whole packet can be received by the serving DA in the current transmission frame. $\mathcal{R}_d^{sjn}$ denotes the average achievable bit rate by a $j$ ranked device located at $d$ and scheduled on the $n^{th}$ sub-channel with $s$ scheduled devices. Using Shannon's channel capacity formula, coverage probabilities given in (4.7), and SIRs given in (4.6), $\mathcal{R}_d^{sjn}$, conditioned on the device is in coverage (i.e., $\Xi_{0j}^{ns} > \tau \in \{\tau_{ah}, \tau_{al}\}$), can be derived as

$$\mathcal{R}_d^{1hn} = E\left[\omega_n \log_2\left(1+\Xi_{0h}^{n1}\right) \middle| \Xi_{0h}^{n1} > \tau_{ah}\right]. \tag{4.20}$$

$$\mathcal{R}_d^{2hn} = E\left[\omega_n \log_2\left(1+\Xi_{0h}^{n2}\right) \middle| \Xi_{0h}^{n2} > \tau_{ah}\right]. \tag{4.21}$$

$$\mathcal{R}_d^{2ln} = E\left[\omega_n \log_2\left(1+\Xi_{0l}^{n2}\right) \middle| \Xi_{0l}^{n2} > \tau_{al} \cap \Xi_{0h}^{n2} > \tau_{ah}\right]. \tag{4.22}$$

Notice that, as the closed form expression for the PDF of the SIR is not available, the expectation terms in of the data rate cannot be solved. However, what the expectation signifies is the fact that, when a device is in the coverage of its serving DA, it achieves a certain average bit rate. This average bit rate is lower bounded by a fixed bit rate allocation given by

$$\mathcal{L}_d^{sjn} = \begin{cases} \omega_n \log_2(1+\tau_{ah}), & s=1 \;\&\; j=h \\ \omega_n \log_2(1+\tau_{ah}), & s=2 \;\&\; j=h \\ \omega_n \log_2(1+\tau_{al}), & s=2 \;\&\; j=l. \end{cases} \tag{4.23}$$

Another possible approximation of $\mathcal{R}_d^{sjn}$ is the unconditional average achievable bit rate which can be derived by removing the condition from the expectation terms, and making use of the definition $E[X] = \int_{t>0} P(X > t)\, dt$, as the logarithm function is non-negative. Accordingly, the unconditional achievable bit rate, $\mathcal{F}_d^{sjn}$, is given by (see appendix B.3)

$$\mathcal{F}_d^{1hn} = \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \tag{4.24}$$

$$\mathcal{F}_d^{2hn} = \int_{t>0} E\left[\exp\{-\mathcal{B} g_{0l}^{n0}\}\right] \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \tag{4.25}$$

$$\mathcal{F}_d^{2ln} = \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \tag{4.26}$$

where $\mathcal{B} = 2^{\frac{t}{\omega_n}} - 1$.

Therefore, for the purpose of this analysis, we assume that the average achievable bit rate by a device is given by

$$\mathcal{R}_d^{sjn} = \max\{\mathcal{L}_d^{sjn}, \mathcal{F}_d^{sjn}\}, \quad \text{for } s \in \{1,2\} \;\&\; j \in \{h,l\}. \tag{4.27}$$

### 4.3.2 DA-BS queue analysis

Similar to the case of a device, for a DA to be able to transmit data to the serving BS, it first needs to access the BS. Let $\Gamma_a$ denote the number of transmission frames needed for a DA to be able to transmit its data successfully to the serving BS, which is modelled as a geometric distribution with success probability $\Theta(\Xi_a^k)$ given by (4.16). Accordingly, the average number of transmission frames needed for a DA to successfully access its serving BS is given by $\bar{\Gamma}_a = 1/\Theta(\Xi_a^k)$. To be able to model the behaviour of packet arrivals and departures at a DA, we focus on the period $\Gamma_a^* = T_f \cdot \lceil \bar{\Gamma}_a \rceil$, which is an approximation for the average access delay in terms of an integer number of transmission frames.

Once a DA accesses the BS, it transmits its data during the relay phase of the current transmission frame. Therefore, during the period of $\Gamma_a^*$, a DA can transmit only for a period of

$T_r < \Gamma_a^*$. Since the transmission rate of a DA depends on the amount of allocated resources and its SIR value at the serving BS, using Shannon's channel capacity formula, SIR given in (4.15), and interference model described in Section-4.2, the average achievable bit rate, $\mathcal{R}_a^k$, conditioned on the DA being covered by the serving BS, by a DA located at $a$ and transmitting its data on the $k^{th}$ sub-channel to the BS located at the origin is given by

$$\mathcal{R}_a^k = E\left[\frac{\Omega_b}{\mathcal{K}}\log_2(1+\Xi_a^k)\,\middle|\,\Xi_a^k > \tau_b\right]. \tag{4.28}$$

Similar to the case of a device, the conditional expectation given in (4.28) cannot be solved as the PDF of $\Xi_a^k$ is not available. Thus, the average achievable bit rate by a DA in coverage is approximated by its lower bound, $\mathcal{L}_a^k = \frac{\Omega_b}{\mathcal{K}}\log_2(1+\tau_b)$, or the unconditional average achievable bit rate, given by

$$\begin{aligned}\mathcal{F}_a^k &= E\left[\frac{\Omega_b}{\mathcal{K}}\log_2(1+\Xi_a^k)\right] \\ &\overset{a}{=} \int_0^\infty \left(2\pi\lambda_a \int_0^\infty y e^{-(\pi\lambda_a y^2)} \mathscr{L}_{I_a}\left\{\mathcal{B}_a y^{\alpha(1-\epsilon)}\right\} dy\right) dt\end{aligned} \tag{4.29}$$

where we use the definition $E[X] = \int_{t>0} P(X > t)\, dt$ to compute the expectation of the logarithmic term with respect to $\Xi_a^k$, and the Laplace term in (a) follows from the DA-BS coverage probability given in (4.16) with $\mathcal{B}_a = 2^{\frac{t\mathcal{K}}{\Omega_b}} - 1$ substituted for $\tau_b$. Accordingly, the average achievable bit rate by a covered DA is given by

$$\mathcal{R}_a^k = \max\{\mathcal{L}_a^k, \mathcal{F}_a^k\}. \tag{4.30}$$

The average number of packets that a DA can transmit during the period $\Gamma_a^*$ is given by

$$\mu_a = \left\lfloor \frac{T_r \mathcal{R}_a^k}{\mathcal{D}} \right\rfloor \tag{4.31}$$

where, just like in (8), the floor function $\lfloor\cdot\rfloor$ is to ensure that a packet is transmitted from a DA if and only if it can be fully received at the BS during the current transmission frame.

Packets arrive at a DA from its scheduled devices in each transmission frame. The average number of packets arriving at a DA depends on the number of associated devices, the probability of successful scheduling on a NOMA sub-channel, and the SIR values of the received signals, and the length of the aggregation phase of each transmission frame. Accordingly, let $\gamma_a^i$ denote the packet arrival rate at a DA during the $i^{th}$ transmission frame, given by

$$\gamma_a^i = \mathcal{N} \cdot \left(P(S^n = 1)\mu_d^{1hn} + P(S^n = 2)(\mu_d^{2hn} + \mu_d^{2ln})\right). \tag{4.32}$$

Following from the preceding analysis, for the period of $\Gamma_a^*$, packet arrivals and transmissions at a DA can be modelled as a G/G/1 queue with packet departure rate given in (4.31), and packet arrival rate given by

$$\gamma_a \approx \gamma_a^i * \lceil \bar{\Gamma}_a \rceil. \tag{4.33}$$

In summary, the proposed NOMA-enabled two-hop architecture can be modelled as a tandem queue with packet arrival rate, $\gamma$, at the first queue (i.e., at a device), arrival process at the second queue (i.e., at a DA) having $\gamma_a$ packet arrival rate, and departure process from the second queue with departure rate of $\mu_a$. Based on the performance measures of the two-stage tandem queues, the expected end-to-end system delay experienced by a packet transmitted by a device located at $d$ to a BS at the origin, going through a DA located at $a$, under the first in first out service rule, is given by [20]

$$\Delta t_{e2e} = \frac{\gamma_a + \mu_a - 2\gamma_d}{\gamma_a \mu_a - (\gamma_a + \mu_a)\gamma_d + (\gamma_d)^2}. \tag{4.34}$$

## 4.4 Energy consumption

We define energy consumption of a network component as the energy consumed by a device or a DA in the process of successfully transmitting/relaying a single data packet of size $\mathcal{W}$ bits towards the serving BS. We model the devices and DAs as wireless transceivers [84]. A transceiver consists of four major energy consuming blocks: transmission block (TX), receiver block (RX), local oscillator block (LO), and power amplification (PA) block. The power consumption of a transceiver when acting as a transmitter, $\mathcal{Q}_T$, and as a receiver, $\mathcal{Q}_R$, are given respectively by

$$\mathcal{Q}_T = \mathcal{Q}_{PA} + \mathcal{Q}_{LO}^i + \mathcal{Q}_{TX}, \tag{4.35}$$

$$\mathcal{Q}_R = \mathcal{M}_a \mathcal{Q}_{LO}^i + \mathcal{Q}_{RX} + \mathbb{1}(S^n = 2)\mathcal{Q}_{SIC} \tag{4.36}$$

where $\mathbb{1}(\cdot)$ is the indicator function to check if a sub-channel has at least two devices scheduled and hence the need for SIC, the rate of power consumption of the RX and TX blocks are denoted by $\mathcal{Q}_{RX}$ and $\mathcal{Q}_{TX}$ respectively; $\mathcal{Q}_{LO}^i \in \{\mathcal{Q}_{LO}^d, \mathcal{Q}_{LO}^a\}$ where $\mathcal{Q}_{LO}^d$ and $\mathcal{Q}_{LO}^a$ are non-negative constant power consumption rate of the LOs of a device and a DA respectively; PA power consumption rate is denoted by $\mathcal{Q}_{PA} \in \{\mathcal{Q}_d^t, \mathcal{Q}_a^t\}$ for a device and a DA respectively, where $\mathcal{Q}_d^t = q_d ||d - a||^\alpha$ and $\mathcal{Q}_a^t = q_a ||a - b||^{\epsilon\,\alpha}$; $\mathcal{Q}_{SIC}$ is the fixed power consumption rate by a DA when performing SIC on a sub-channel with more than one scheduled device.

### 4.4.1 Energy consumption of a device

Energy consumption of a device to transmit a data packet is divided into two parts: energy consumed to successfully access the serving DA, and energy consumed to successfully transmit the data packet. We focus on a device with a single data packet in its queue.

Every time the device is scheduled on a NOMA sub-channel, it attempts to access its serving DA. If the device fails to access, it gives up its scheduled sub-channel and waits until the next time it is scheduled. The device attempts access $\lceil \bar{\Gamma}_d \rceil$ times until it gains access. Out of those attempts, the device is scheduled $W$ times, which is a binomial random variable with parameters $\lceil \bar{\Gamma}_d \rceil$ and $\bar{\Lambda}_s$. Considering that every time a scheduled device attempts to transmit its data packet and fails, it consumes a fixed amount of energy denoted by $\mathcal{E}_d^f$, the access energy consumption of a device is $\mathcal{E}_{d,1} = \lceil \bar{\Gamma}_d \rceil \cdot \bar{\Lambda}_s \cdot \mathcal{E}_d^f$.

Once the device accesses the DA, it successfully transmits its data packet over the allocated sub-channel. Let $\Delta t_x$ denote the average time a device takes to transmit a single data packet of size $\mathcal{W}$ to its serving DA, given by

$$\Delta t_x = \frac{\mathcal{D}}{P(S^n = 1)\mathcal{R}_d^{1hn} + P(S^n = 2)(\mathcal{R}_d^{2hn} + \mathcal{R}_d^{2ln})}. \tag{4.37}$$

The associated average transmission energy consumption is

$$\mathcal{E}_{d,2} = \Delta t_x \left( E[Q_d^t] + Q_{LO}^d + Q_{TX} \right) \tag{4.38}$$

where $E[Q_d^t] = \int_0^\infty 2q_d \pi \lambda_a r^{(\alpha+1)} e^{-\pi \lambda_a r^2} dr$, as $Q_d^t$ is a function of the distance between the device and the DA having a Rayleigh distribution with parameter $\lambda_d$.

Thus, the total average energy consumption of a device to transmit a single data packet of size $\mathcal{W}$ bits in the proposed NOMA-enabled two hop network is $\mathcal{E}_d = \mathcal{E}_{d,1} + \mathcal{E}_{d,2}$.

### 4.4.2 Energy consumption of a DA

Different from an IoT device, a DA consumes energy in two forms: as a receiver while receiving data from its cluster members, and as a transmitter while relaying the aggregate data packets to the BS. We focus on a DA located at $a$ and associated with the BS at the origin.

The total time for a device to transmit a single data packet of size $\mathcal{W}$ bits to the DA is $\Delta t_x$. The DA needs to stay awake for this time period to successfully receive the data packet. Thus, let $\mathcal{E}_{a,1}$ denote the amount of energy that the DA consumes to receive a single data packet, given by

$$\mathcal{E}_{a,1} = \Delta t_x \left( Q_{LO}^a + Q_{RX} + P(S^n = 2) Q_{SIC} \right). \tag{4.39}$$

The DA then attempts to relay the data packet and in the process consumes energy in two forms: energy consumed to successfully access the BS, and energy consumed to successfully transmit the packet. The average number of transmission frames needed for the DA to successfully access the BS is $\lceil \bar{\Gamma}_a \rceil$. Consider that a DA consumes a fixed amount of power, $\mathcal{E}_a^f$, in every access attempt. Then, the average access energy consumption is $\mathcal{E}_{a,2} = \mathcal{E}_a^f \lceil \bar{\Gamma}_a \rceil$.

Once the DA successfully accesses the BS, it transmits the data packet. The transmission energy consumption for the data packet is given by

$$\mathcal{E}_{a,3} = \Delta t_a \left( E_Y [Q_a^t] + Q_{LO}^a + Q_{TX} \right) \tag{4.40}$$

where $E_Y [Q_a^t] = \int_0^\infty 2 q_a \pi \lambda_b y^{(\epsilon \alpha + 1)} e^{-\pi \lambda_b y^2} \, dy$, where we used $y$ to denote the random transmission distance between the DA and the BS which follows a Rayleigh distribution (given in (3.4)) with parameter $\lambda_b$, and $\Delta t_a = \mathcal{D} / \mathcal{R}_k^a$.

Therefore, the total energy consumption of a DA to relay a single data packet of size $\mathcal{W}$ bits in the proposed NOMA-enabled two hop network is $\mathcal{E}_a = \mathcal{E}_{a,1} + \mathcal{E}_{a,2} + \mathcal{E}_{a,3}$.

## 4.5 Numerical results and discussion

We evaluate the developed models via computer simulations for the scheduling probability, device-DA coverage probability, total coverage probability, system delay, device energy consumption, and DA energy consumption, as functions of DA density $\lambda_a$ and thresholds $\tau_{ah}$, $\tau_{al}$ and $\tau_b$. The numerical results are validated by means of independent system level simulations. Table 4.1 lists parameter values used.

Table 4.1: Simulation parameter values

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| $\mathcal{W}$ | 10 bits/packet | $\mathcal{E}_d^f$ | 0.05 mJ |
| $\mathcal{E}_a^f$ | 0.05 mJ | $\mathcal{N}$ | 10 |
| $Q_{SIC}$ | 0.1 mW | $Q_{LO}^d$ ($Q_{LO}^a$) | 0.1 mW |
| $Q_{RX}$ ($Q_{TX}$) | 0.1 mW | $q_d$ ($q_a$) | 1 |
| $T_f$ ,($T_a$,$T_r$) | 10 ms (5 ms, 5 ms) | $\alpha$ | 4 |
| $\gamma$ | 10 packets/s | $\epsilon$ | 0.8 |
| $\lambda_d$ | 1k dev/km$^2$ | $\lambda_b$ | 2 BS/km$^2$ |
| $\omega_b$ | 50 kHz | $\omega_n$ | 5 kHz |

Simulations are performed based on the system model in Section 3.2 using MATLAB for a 200 $km$ x 200 $km$ 2D plane. Each result is an average of 1000 independent Monte Carlo simulation runs. Locations of BSs, DAs, and devices are randomly generated based on their deployment densities. The wireless channel is simulated with the 3GPP non-line-of-sight (NLOS) path loss model given in [25]. To simulate Rayleigh fading, random channel gains are generated according

to an exponential distribution with unity mean and incorporated into the received signal power of all transmissions. Maximum received signal strength association rule is employed. Through extensive simulations, the derived models are shown to be functional and accurate as evident by the close match between the numerical and simulation results.

Figure 4.3 shows the scheduling probability of the devices for NOMA and OMA configurations versus DA density $\lambda_a$ between 95 and 435 DAs/$km^2$. For both NOMA and OMA, the scheduling probability increases as DA density increases. As expected, the scheduling probability is higher for the NOMA as compared to OMA, thanks to its ability to allow multiplexing of more than one device on the same sub-channel.

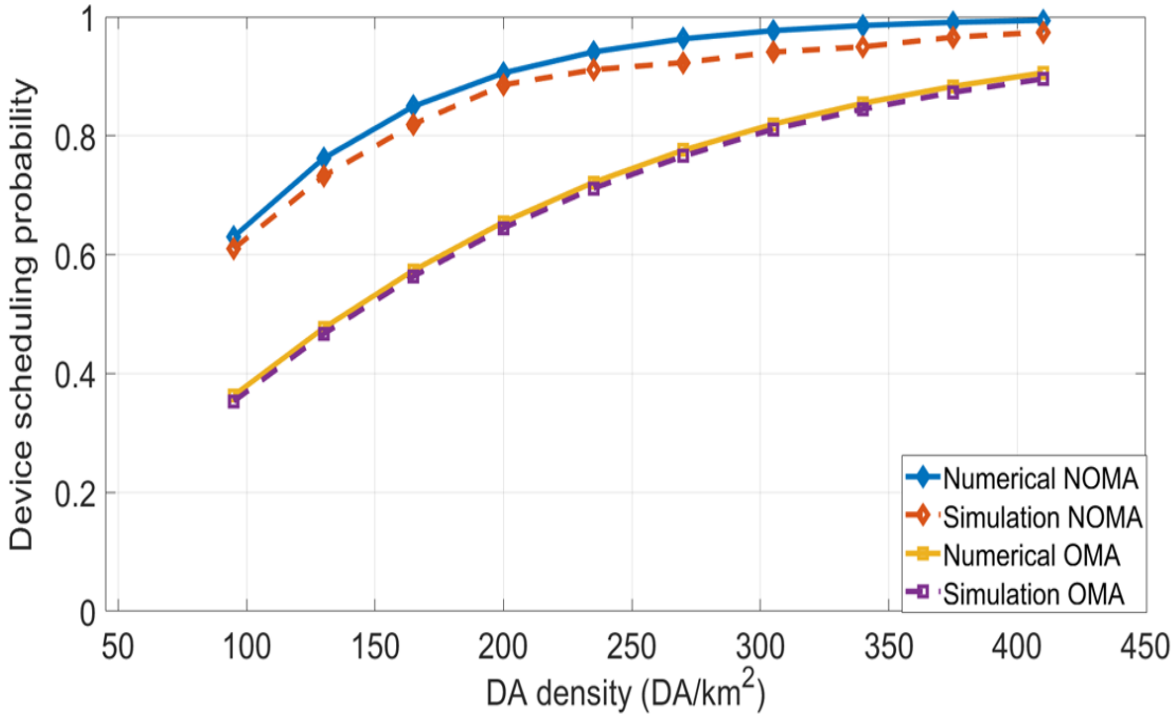


Figure 4.3: Scheduling probability of the devices as a function of DA density ($\lambda_a$)

Figure 4.4 shows device-DA coverage probability ($C_d$) and total coverage probability ($C_{tot}$) versus DA density $\lambda_a$, and network thresholds $\tau_{ah} = \tau_{al} = \tau_b = \tau$. The $C_{tot}$ values are slightly lower than those of $C_d$, as $C_{tot}$ takes into account the performance of both hops. However, it is clear that the device-DA coverage is more dominant in our system, which is expected due to the extensive interference among the devices and the high contention over the limited number of in-cluster sub-channels. As both coverage probabilities follow the same trends, we focus our analysis only on $C_d$. For both NOMA and OMA, as $\lambda_a$ increases from 95 DAs/$km^2$ to 435 DAs/$km^2$, $C_d$ increases to a maximum and then drops in a concave manner. This concave pattern can be attributed to the fact that, as $\lambda_a$ increases, the number of interfering devices on a sub-channel increases, leading to a decrease in the probability of successful transmission

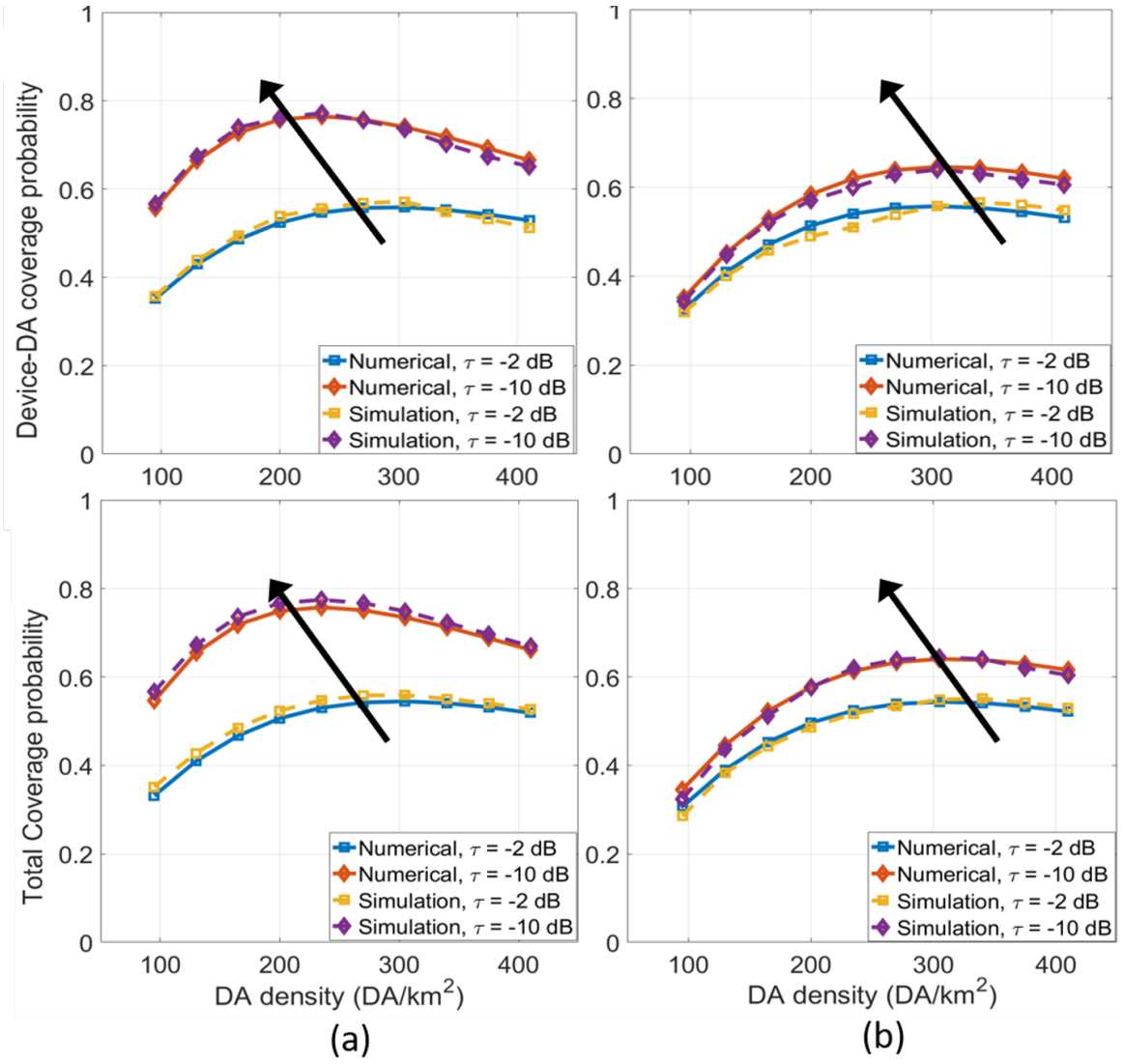


Figure 4.4: Device-DA coverage probability (top) and total coverage probability (bottom) as a function of DA density ($\lambda_a$) and network thresholds ($\tau$); (a) NOMA; (b) OMA

from a device to the serving DA. For NOMA, the maximums occur at densities of 305 DA/$km^2$, for $\tau = -2\ dB$, and 235 DA/$km^2$, for $\tau = -10\ dB$, whereas in the case of OMA, the maximum for both $\tau$ settings happens at 305 DA/$km^2$. Accordingly, we can conclude that there is a $\lambda_a$ that maximizes $C_d$, given a fixed density of devices for both NOMA and OMA. Also, a lesser number of clusters is needed in the case of NOMA to maximize performance. On the other hand, for both configurations, the coverage probability improves as $\tau$ drops, and the improvement for NOMA is more significant.

Figure 4.5 shows the total system delay versus DA density for multiple network thresholds for both NOMA and OMA. NOMA has a lower average system delay as compared to OMA. The difference is further aggravated at higher DA densities. Thanks to its ability to multiplex multiple devices on the sub-channels, NOMA not only allows scheduling a larger number of

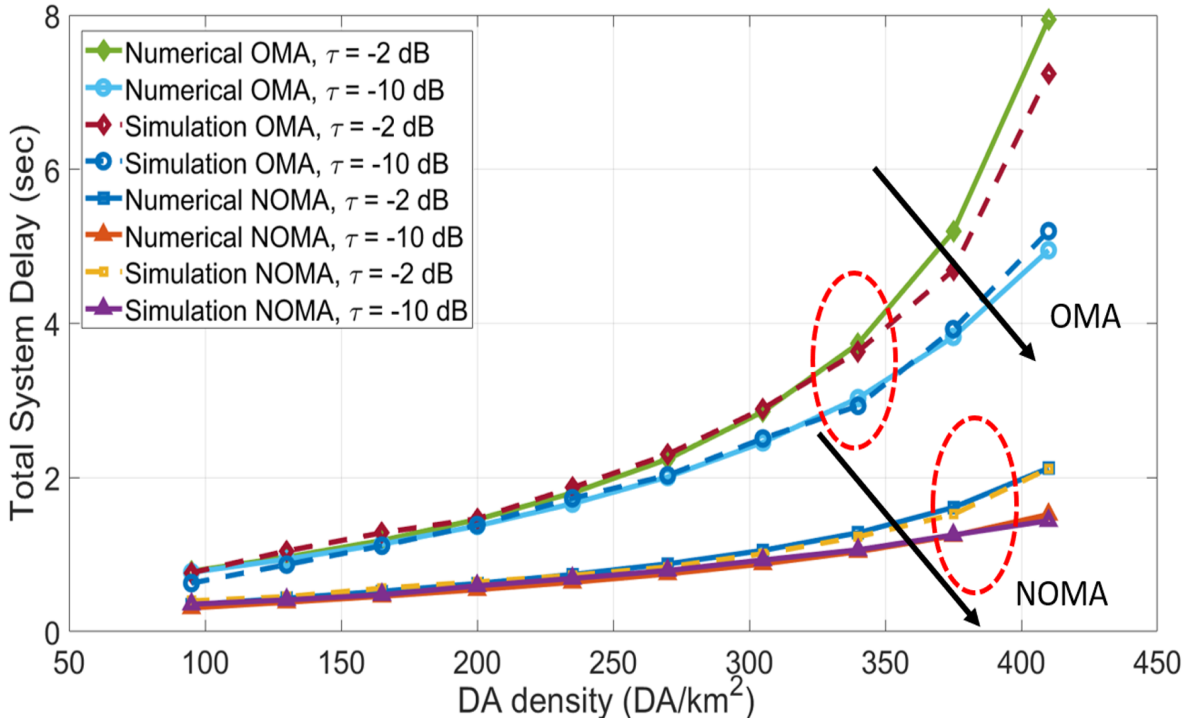


Figure 4.5: Total system delay as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

devices per transmission frame, but also provides a better coverage probability, especially for lower network threshold settings. Accordingly, packet departure rate from a device is higher for NOMA, leading to shorter queuing delays and overall improvement in system delay. On the other hand, the delay increases as $\lambda_a$ increases as more DAs share a limited bandwidth, leading to a lower achievable bit rate in packet departure from a DA.

Figure 4.6 shows the average energy consumption of device to transmit a packet of size $\mathcal{W}$ bits to the serving DA for both NOMA and OMA configurations versus $\lambda_a$ and $\tau$. For both NOMA and OMA, energy consumption increases as $\lambda_a$ and $\tau$ values increase except for NOMA at $\tau = -2dB$. The higher the $\tau$, the higher the number of transmission frames required before a device can access the serving DA. At higher $\lambda_a$ values, interference among devices on a sub-channel increases, leading to a higher number of failed attempts and in turn higher levels of energy wastage. For NOMA at $\tau = -2dB$, the impact of $\tau$ is more dominant than the effect of $\lambda_a$, because devices experience higher interference than that under OMA. Thus, as $\tau$ increases, the SIR performance decreases at a faster rate for NOMA. Further, for the same $\tau$ setting, devices under NOMA configuration consume more energy to transmit a data packet compared to under OMA for all DA densities, because the average achievable bit rate by a device is lower in the case of NOMA due to intra-cluster interference. A lower bit rate results in a longer time that a device takes to transmit a single data packet, leading to higher device energy consumption.

The average energy consumed by a DA to relay a data packet of size $\mathcal{W}$ bits, can be of

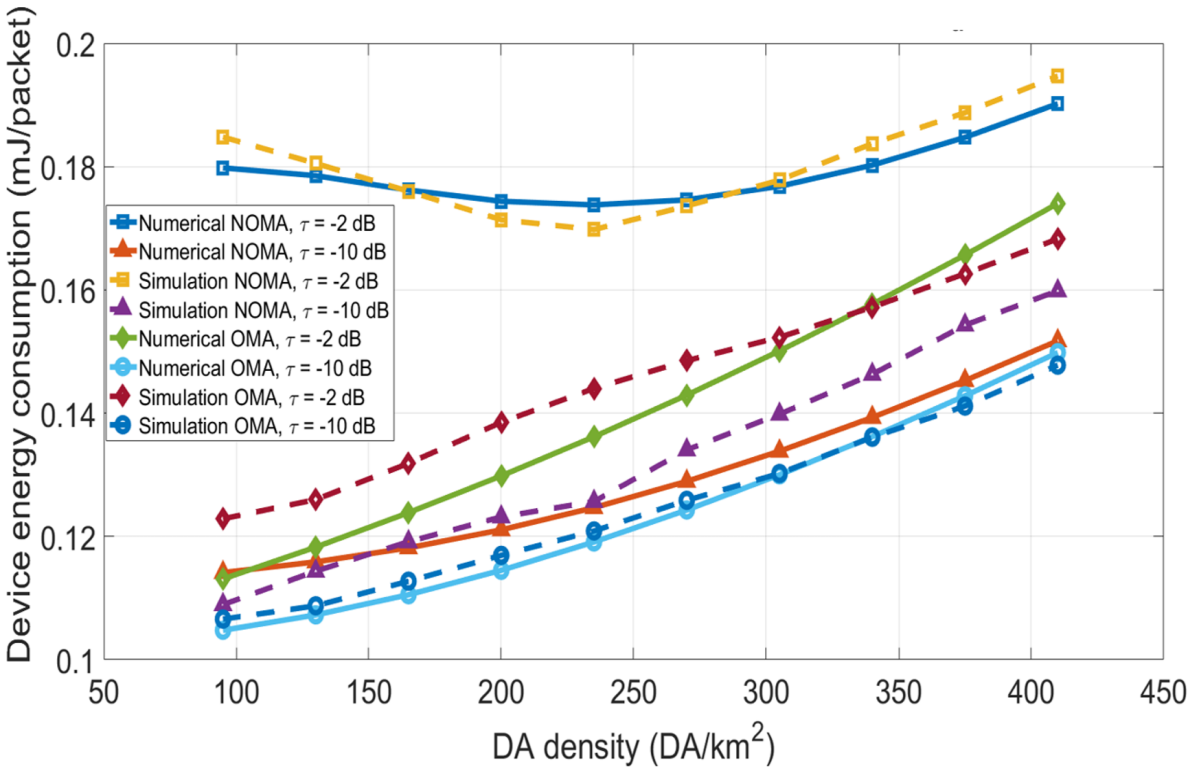


Figure 4.6: Average device energy consumption as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

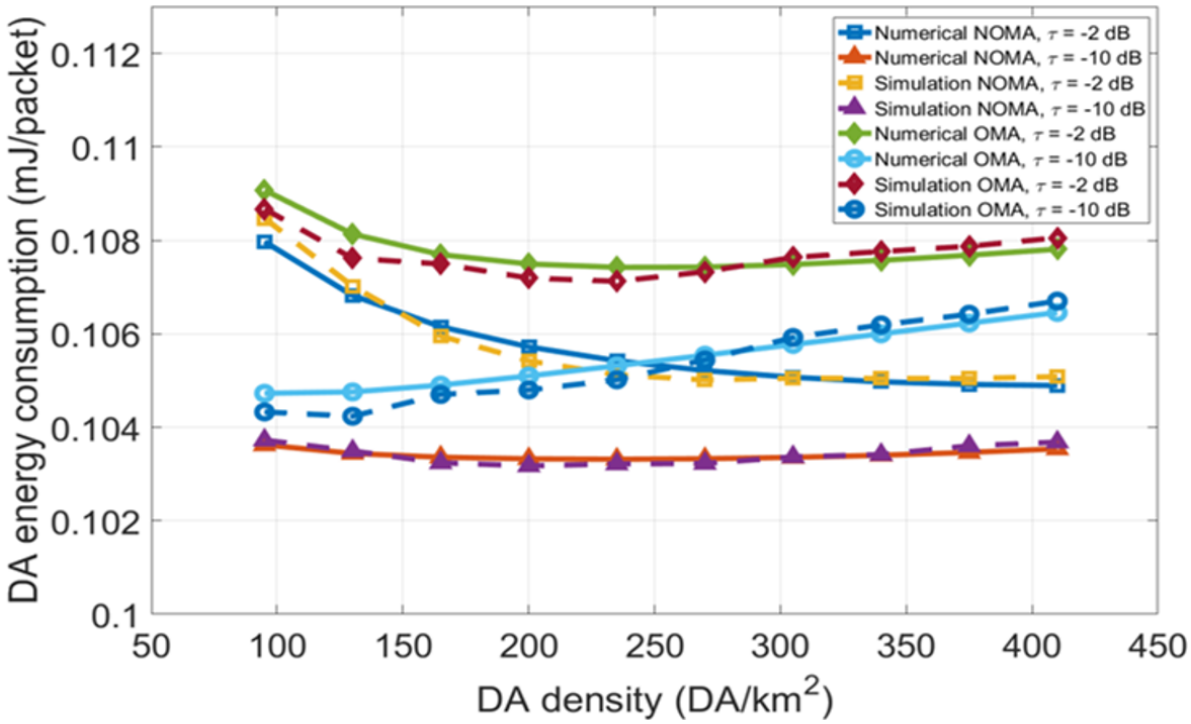


Figure 4.7: Average DA energy consumption as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

a concern, especially for DAs selected from the IoT devices. As shown in Figure 4.7, for the same $\tau$ setting, DAs consume less amount of energy under NOMA. As $\lambda_a$ increases, DA energy consumption either decreases (in the case of $\tau = -2\ dB$), or is almost stable (in the case of $\tau = -10\ dB$). On the other hand, for OMA, as DA density increases, DA energy consumption increases for all $\tau$. Overall, NOMA is more energy efficient for the DAs.

Based on the preceding results, for a two-hop NOMA enabled data aggregation architecture, the following general conclusions can be made: i) NOMA for in-cluster transmissions can improve the scheduling probability of the devices in the case of limited transmission resources; ii) NOMA can improve the total coverage probability of the two-hop network and in turn increase the number of supported devices per transmission frame; iii) with NOMA, end-to-end delay can be improved with the proper choice of network parameters such as DA density and amount of available resources; and iv) with a proper interference mitigation mechanism for the in-cluster NOMA transmissions and the appropriate choice of the DA density, NOMA can improve the overall energy efficiency of both DAs and devices. Based on the results presented in this study, our proposed NOMA enabled architecture has many potentials and further investigations are needed to investigate the energy and delay optimality design of two-hop NOMA-enabled transmission architecture for future massive cellular IoT applications.

## 4.6 Summary

In this Chapter, we propose a novel two-hop NOMA enabled data aggregation architecture to enable massive cellular IoT applications. Concepts and techniques from stochastic geometry and queuing theory are jointly exploited to derive tractable models for various network performance measures including scheduling probability, coverage probability, system delay, and energy consumption. We abstract the network topology by spatially modeling the locations of the BSs, DAs and active devices using three independent homogeneous PPPs, and characterize intra- and inter-cluster interference components experienced by the devices and characterize the inter-cellular interference experienced by the DAs. All derived models are corroborated via Monte Carlo simulations. Numerical results demonstrate that, the DA density and network thresholds highly impact network performance. In comparison with the traditional two-hop OMA based architecture, the proposed NOMA architecture provides better scheduling and coverage probability, supports a larger number of devices per transmission frame, has a lower average end-to-end system delay, and improves the energy consumption of the devices and DAs.

# Chapter 5

# Access Point Association in Uplink Two-Hop cellular IoT Networks with Data Aggregators

As was shown in Chapter 4, node clustering and data aggregation help extend the coverage of cellular networks and increase the number of supported devices, while meeting the various service quality requirements and reducing energy consumption, making them suitable for enabling future massive cellular IoT applications. They also help alleviate the congestion at the PRACH by reducing the number of devices contending over the limited number of RACH preambles. Nonetheless, in the system model of Chapter 4, we assume that all devices transmit their data in two-hop fashion via a middle DA node. In that scenario, the coverage areas of DAs form a Voronoi tessellation. This assumption is plausible when the number of DAs is very large such that they can provide full network coverage. However, in more realistic settings, as the DAs have a limited coverage area and their coverage depends on the channel conditions and path-loss attenuation, the continuous Voronoi tessellation can be somewhat inaccurate. This creates a network with heterogeneous coverage areas. Some areas may be covered only by BSs, while other might be covered by a BS and one or more DAs. In such scenario, to which AP a device associates is critical and impacts the overall performance of the network, especially when DAs share the same cellular resources with the devices.

In this Chapter, we consider a more realistic scenario where we overlay the cellular network with a layer of cellular DAs, such that the network provides both single and two-hop routes. As DAs share radio resources with single-hop devices, a dependency between the two routes is present. We recognize that the proper design of the DA enabled network becomes critical for cost effectiveness and efficient radio resource utilization. We tackle this problem from the viewpoint of user association, where active devices need to decide on which AP to associate with and use to transfer their data to the core network while satisfying their QoS requirements. To

that end, in this Chapter, we consider minimum required data rate as the QoS requirements. The user association problem is formulated as a joint AP association, resources utilization, and energy efficient communication optimization problem that takes into account various networking factors such as the number of devices, number of DAs, number of available resource units, interference, transmission power limitation of the devices, DA transmission performance, and channel conditions. The objective is to show the usefulness of data aggregation and shed light on the importance of network design when the number of devices is massive. We propose a coalition game theory based algorithm, *PAUSE*, to transform the optimization problem into a simpler form that can be successfully solved in polynomial time. Different network scenarios are simulated to showcase the effectiveness of *PAUSE* and to draw observations on cost effective network design with DAs.

The rest of this chapter is organized as follows. System model is presented in Section 5.1. Section 5.2 details the problem formulation and presents the general multi-objective optimization problem. In Sections 5.3 and 5.4, we describe the proposed heuristic algorithm to efficiently solve the joint optimization problem, and discuss its convergence and complexity. In Section 5.5, we evaluate the performance of the algorithm by means of simulations. We conclude this study in Section 5.6.

## 5.1 System model

### 5.1.1 Physical network

Consider an uplink network made up of a single layer of BSs (Figure 5.1). BS locations form a point Poisson process (PPP) $\Phi_b = b_1, b_2, ...$ with density $\lambda_b$, where $b_i$ denotes the location of the $i^{th}$ BS in the network. Let $\mathcal{B} = |\Phi_b|$ denote the cardinality of set $\Phi_b$, i.e., the number of BSs in the network. The network is overlaid with a layer of DAs uniformly distributed over the network coverage. DA locations form a PPP $\Phi_a = a_1, a_2, ...$ with density $\lambda_a > \lambda_b$, where $a_i$ denotes the location of the $i^{th}$ DA. Let $\mathcal{A} = |\Phi_a|$ denote the cardinality of set $\Phi_a$. Coverage area of a DA is much smaller than that of a BS. We approximate the coverage area of a DA by a disk of radius $\Lambda_a$. In the following, we use AP to denote both BS and DA wherever there is no ambiguity.

The network supports a large number, $\mathcal{D}$, of low mobility and battery powered IoT devices. Let $\Phi_d = \{d_1, d_2, ..., d_{\mathcal{D}}\}$ denote the location set of the devices, where $d_i$ denotes the location of the $i^{th}$ device. A device can associate with only one AP at a time. A device may be located in a single covered area (i.e., area covered only by a BS), or in a double covered area (i.e., area covered by a BS and one or more DA). A device located in a double covered area has the choice

of associating with the BS or the DA that best meets its QoS requirements and minimizes its transmission power.

Denote the set of devices associated with BS $j$ as $\Upsilon_{b_j} = \{d^1_{b_j}, d^2_{b_j}, ..., d^k_{b_j}\}$, where $d^i_{b_j}$ denotes the location of the $i^{th}$ device associated with BS $b_j \in \Phi_b$ . Similarly, let $\Upsilon_{a_h} = \{d^1_{a_h}, d^2_{a_h}, ..., d^g_{a_h}\}$ denote the location set of the devices associated with the $h^{th}$ DA, where $d^i_{a_h}$ denote the location of the $i^{th}$ device associated with DA $a_h \in \Phi_a$. Finally, let $\Upsilon_b = \{\Upsilon_{b_1}, \Upsilon_{b_2}, ..., \Upsilon_{b_\mathcal{B}}\}$ and $\Upsilon_a = \{\Upsilon_{a_1}, \Upsilon_{a_2}, ..., \Upsilon_{a_\mathcal{A}}\}$ denote the super location sets of the assigned devices to all BSs and DAs respectively. Notice that a device may not be associated in a transmission frame, thus $|\Upsilon_b| + |\Upsilon_a| \leq \mathcal{D}$. Also, as a device is allowed to associate with only a single AP at a time, $\Upsilon_k \cap \Upsilon_j = \emptyset$, for $k$ & $j \in \{\Phi_b, \Phi_a\}$.

We use binary indication variables $x_{d_i,b_j}$ and $y_{d_i,a_h}$ to indicate the association status of the $i^{th}$ device with the $j^{th}$ BS and the $h^{th}$ DA respectively. The variables equal 1 if the device is associated with that particular AP, and 0 otherwise. A device cannot have both $x_{d_i,b_j}$ and $y_{d_i,a_h}$ equal to 1 as the device can only associate with one AP at a time. A device is considered associated to an AP if and only if the AP is able to provide the necessary QoS level required by the device.

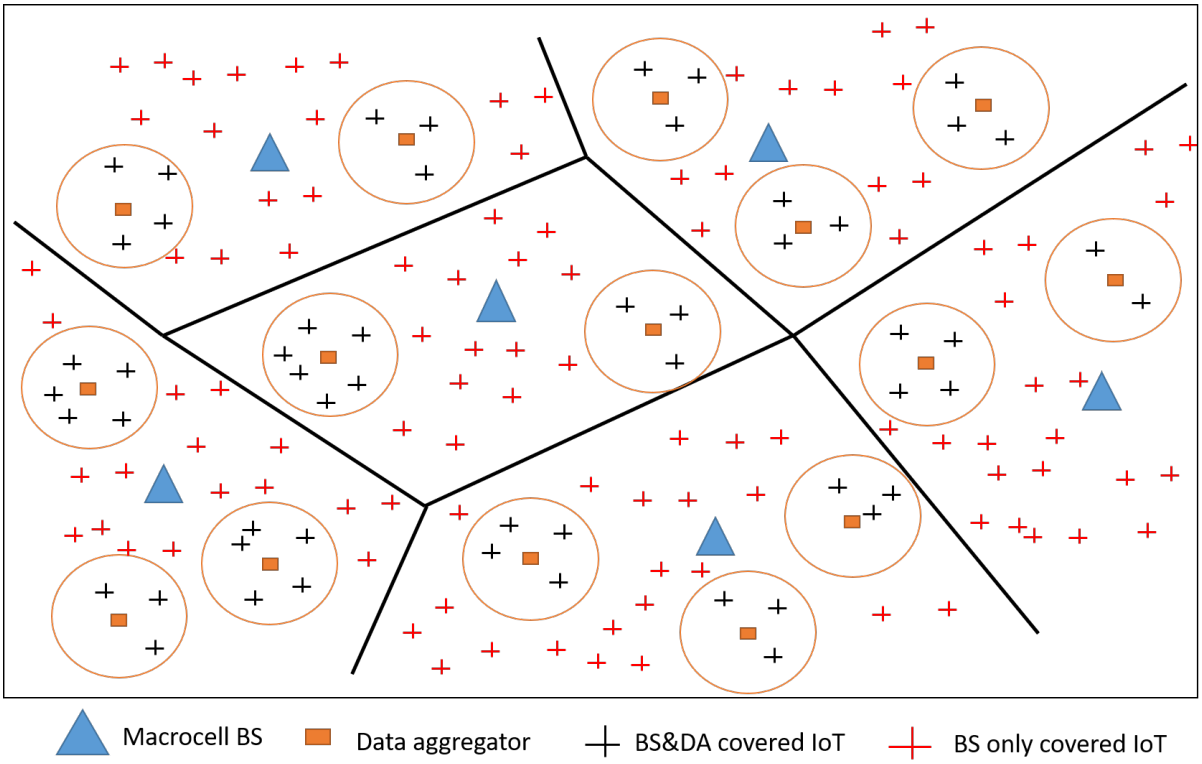


Figure 5.1: An illustration of BS, DA, and IoT device locations as well as the Voronoi tessellation formed by the coverage of the BSs and the circular coverage of the DAs.

### 5.1.2 Wireless transmission model

The uplink radio spectrum at the BS is divided into two orthogonal sub-bands, one dedicated for DA transmissions, $C_a$, and $C_d^1$ for directly connected devices. Sub-band $C_a$ is divided into $L = \{1, 2, ..., \mathcal{L}\}$ orthogonal channels of equal bandwidth $\omega_u$, referred to as resource units (RUs). A DA can occupy a single RU at a time, and an RU at a BS can be allocated to one DA. DAs from different BSs may be scheduled on the same RU and, thus, DA transmissions may suffer from inter-cell interference. We assume a fully loaded network in which all RUs at each BS has a DA scheduled and that DAs do not change their RU assignments.

Sub-band $C_d^1$ is divided into, $K = \{1, 2, ..., \mathcal{K}\}$, equal bandwidth, $\omega_b$, channels referred to as resource blocks (RBs). Depending on the QoS requirements, interference level, location, and transmission power, a single device may be allocated one or more RBs to accommodate its needs. Each device connected to a BS can be scheduled on one RB at a time; however, multiple devices can be scheduled on the same RB while connecting to different BSs. Thus, directly connected devices may experience inter-cellular interference.

Devices that transmit in two-hop fashion via DAs, share a device-DA sub-band, denoted by $C_d^2$, that is orthogonal to both bands $C_a$, and $C_d^1$. Sub-band $C_d^2$ consists of $N = \{1, 2, ..., \mathcal{N}\}$ resource channels (RCs) of equal bandwidth $\omega_c$. Similar to BS-directly connected devices on the RBs, a device connected to a DA on $C_d^2$ may be assigned one or more RCs depending on its QoS requirements, signal to interference ratio (SIR), and resource availability. An RC can only be used by one device connecting to the same DA, yet devices can be assigned the same RC at different DAs. Thus, two-hop connecting devices experience only inter-cluster interference.

Signal transmissions experience propagation attenuation according to a general power-law path-loss model. The signal power decays at rate $D^{-\alpha}$, where $D$ is the propagation distance and $\alpha$ is the path-loss exponent. All transmissions suffer from Rayleigh fading that introduces random instantaneous power gain, $g$, following an exponential distribution with unity mean (i.e. $g \sim \exp\{1\}$). The channel power gains are distance independent, independent of each other and identically distributed (i.i.d.). The network capacity is interference limited.

DAs are assumed to have access to infinite power supply and hence they employ full power inversion to compensate for path-loss attenuation. Let $q_a$ denote the nominal transmission power of any DA in the network. We focus on a DA located at $a_i$ and associated with the BS located at $b_j$. Accordingly, the initial transmission power of the DA after full power inversion to compensate for the path-loss attenuation is given by $q_{a_i} = q_a ||a_i - b_j||^\alpha$. Consequently, the received power at the BS from the DA is given by $q_{a_i}^r = q_a g$, which eliminates the impact of path-loss attenuation. This assumption is made for simplicity of analysis as it reduces the

complexity due to distance-based path-loss attenuation. In contrast, devices are assumed to be battery powered and thus there is an upper limit on the maximum transmission power that can be used by any device. The transmission power of each device is to be allocated based on the optimization problem that best provides the required QoS satisfaction at the minimum energy consumption. Accordingly, let $q_{d_i} \in \{q^k_{d_i,b_j}, q^n_{d_i,a_h}\}$ denote the transmit power of the $i^{th}$ device when transmitting to AP $\in \{b_j, a_h\}$ on the uplink scheduled resource $x \in \{k, n\}$, for $b_j \in \Phi_b$, $a_h \in \Phi_a$, $k \in K$ and $n \in N$. Let the maximum transmit power of a device be denoted by $Q_{max}$ such that $0 \leq q_{d_i} \leq Q_{max}$.

### 5.1.3 One-hop instantaneous device data rate

Devices of one-hop transmission share $K = \{1, 2, ..., \mathcal{K}\}$ orthogonal RBs, where $k \in K$ is the RB index. Let the channel gain on the $k^{th}$ RB between the $i^{th}$ device (associated with the $j^{th}$ BS) and the $h^{th}$ BS be denoted by $g^k_{d_{i(j)},b_h}$. Also, let $q^k_{d_i,b_j} \geq 0$ and $s^k_{d_i,b_j} \in \{0, 1\}$ be the transmission power level and the binary RB assignment variables of the $i^{th}$ device from the $j^{th}$ BS on the $k^{th}$ RB. As a device can be assigned multiple RBs at the same time, $\mathbf{Q}_{d_i,b_j} = [q^1_{d_i,b_j}, ..., q^K_{d_i,b_j}]$ and $S_{d_i,b_j} = [s^1_{d_i,b_j}, ..., s^K_{d_i,b_j}]$ are vectors of the overall transmission power and RB assignments of the $i^{th}$ device from the $j^{th}$ BS on all $K$ RBs. Accordingly, we have super sets $\mathbf{Q}_b = [\mathbf{Q}_{d_1,b_1}, \mathbf{Q}_{d_2,b_1}, ..., \mathbf{Q}_{d_{|\Upsilon_{b_1}|},b_1}, ..., \mathbf{Q}_{d_{|\Upsilon_{b_j}|},b_j}]$ and $S = [S_{d_1,b_1}, S_{d_2,b_1}, ..., S_{d_{|\Upsilon_{b_1}|},b_1}, ..., S_{d_{|\Upsilon_{b_j}|},b_j}]$ as the collection of the overall transmission power and RB assignment vectors for all BS-connected devices. The achievable data rate by the $i^{th}$ device from the $j^{th}$ BS on the $k^{th}$ RB, using Shannon's channel capacity formula, is

$$R^k_{d_i,b_j} = \omega_b \log_2 \left( 1 + \frac{q^k_{d_i,b_j} ||d_i - b_j||^{-\alpha} g^k_{d_{i(j)},b_j}}{\sum\limits_{b_h \in \Phi_b / b_j} \sum\limits_{d_m \in \Upsilon_{b_h}} s^k_{d_m,b_h} q^k_{d_m,b_h} ||d_m - b_j||^{-\alpha} g^k_{d_{m(h)},b_j}} \right)$$

where $X/x_i$ means all items in set $X$ excluding item $x_i$.

### 5.1.4 Two-hop instantaneous data rate

Different from the one-hop transmission, in the two-hop route, a device first transmits data to its serving DA which then relays the data to the serving BS. The achievable data rate using the two-hop route is the minimum of the achievable rates of the two links. For the first link (device to DA), the same analysis from Sub-section 5.1.3 can be applied with notational variation as follows. As devices connected to DAs share $N = \{1, 2, ..., \mathcal{N}\}$ RCs, where $n \in N$ indicates the RC index, let the channel power gain on the $n^{th}$ RC from the $j^{th}$ DA between the $i^{th}$ device and the $h^{th}$ DA be denoted by $g^n_{d_{i(j)},a_h}$. Also, let $q^n_{d_i,a_j} \geq 0$ and $v^n_{d_i,a_j} \in \{0, 1\}$ be the transmission power level and the binary RC assignment variables of the $i^{th}$ device from the $j^{th}$ DA on

the $n^{th}$ RC. Further, vectors $\mathbf{Q}_{d_i,a_j} = [q^1_{d_i,a_j}, ..., q^N_{d_i,a_j}]$ and $V_{d_i,a_j} = [v^1_{d_i,a_j}, ..., v^N_{d_i,a_j}]$ indicates the overall transmission power and RC assignments of the $i^{th}$ device from the $j^{th}$ DA on all $N$ RCs. Accordingly, we have super sets $\mathbf{Q}_a = [\mathbf{Q}_{d_1,a_1}, \mathbf{Q}_{d_2,a_1}, ..., \mathbf{Q}_{d_{|\Upsilon_{a_1}|},a_1}, ..., \mathbf{Q}_{d_{|\Upsilon_{a_j}|},a_j}]$ and $V = [V_{d_1,a_1}, V_{d_2,a_1}, ..., V_{d_{|\Upsilon_{a_1}|},a_1}, ..., V_{d_{|\Upsilon_{a_j}|},a_j}]$ as the collection of the overall transmission power and RC assignment vectors for all DA-connected devices in the network. The achievable data rate by the $i^{th}$ device on the $n^{th}$ RC from the $j^{th}$ DA, using Shannon's channel capacity formula, is

$$R^n_{d_i,a_j} = \omega_c \log_2\left(1 + \frac{q^n_{d_i,a_j}||d_i - a_j||^{-\alpha} g^n_{d_{i(j)},a_j}}{\sum\limits_{a_h \in \Phi_a/a_j} \sum\limits_{d_m \in \Upsilon_{a_h}} v^n_{d_m,a_h} q^n_{d_m,a_h} ||d_m - a_j||^{-\alpha} g^n_{d_{m(h)},a_j}}\right).$$

As for the second link (DA to BS), DAs share $L = \{1, 2, ..., \mathcal{L}\}$ of RUs when relaying their aggregated data packets. As we consider a fully loaded network of DAs such that there is a DA scheduled on every RU at every BS, let $\Phi^l_a = \{a^l_0, a^l_1, a^l_2, ..., a^l_i\}$ denote the location set of inter-cell interfering DAs on the $l^{th}$ RU, where $a^l_j$ is the location of the DA scheduled on the $l^{th}$ RU of the $j^{th}$ BS, for $j \in \Phi_b$. Further, let $g^l_{a_j,k}$ denote the channel power gain on the $l^{th}$ RU from the $j^{th}$ BS between the inter-cell interfering DA and the $k^{th}$ BS. We focus on a typical DA located at $a^l_0$ and associated with the BS located at the origin. Using Shannon's channel capacity formula, the achievable data rate is given by

$$\mathcal{R}^l_{a_0} = \omega_u \log_2\left(1 + \frac{q_a g^l_{a_0,0}}{\sum\limits_{a^l_j \in \Phi^l_a/a^l_0} q_a ||a^l_j - b_j||^{\alpha} ||a^l_j - b_0||^{-\alpha} g^l_{a_j,0}}\right).$$

As a result, the achievable data rate by the $i^{th}$ device from the $h^{th}$ DA via two-hop route, going through the DA located at $a_h$ towards the serving BS, is (See Appendix C.1)

$$\mathcal{R}_{d_i,2} = min\{\mathcal{R}_{a_h}, \sum_{n}^{\mathcal{N}} v^n_{d_i,a_h} R^n_{d_i,a_j}\} \tag{5.1}$$

where superscript $l$ is omitted from $\mathcal{R}_{a_h}$ under the assumption of static RU assignment to the DAs.

## 5.2 Problem formulation

Devices have QoS requirements that they thrive to achieve. They also want to minimize their energy consumption as they are battery powered devices. On the other hand, due to the limited resources available at the APs, it is crucial to maximize the resource utilization in the system such that the number of satisfied users is maximized. In what follows, we formulate this problem as an

optimization problem to maximize the utility of the network - defined as the maximization of the number of successfully served devices given their data rate requirements and limited resources - while minimizing the energy consumption of the devices - defined as the minimization of the summation the transmit power of all served devices. The outcome of the optimization problem is a set of device-AP association matrices, a vector containing the transmission power assignments of each served device, and a set of RB and RC assignment matrices for resource allocation. For brevity, we write $\mathbf{X} = [x_{d_i,b_j}]$, $\mathbf{Y} = [y_{d_i,a_h}]$, $\mathbf{V} = [v^n_{d_i,a_h}]$, $\mathbf{S} = [s^k_{d_i,b_j}]$, $\mathbf{Q}_b = [q^k_{d_i,b_j}]$, and $\mathbf{Q}_a = [q^n_{d_i,a_h}]$ as BS association matrix, DA association matrix, RC assignment matrix, RB assignment matrix, and transmit power vectors to BSs and DAs respectively. This problem is referred to as the joint AP association, power control, and resource allocation problem, formulated as shown in problems **P1-1** and **P1-2**.

In **P1**, $(C1)$ is to guarantee that a device is assigned service if it is able to obtain an average date rate greater than or equal to its minimum average rate requirement $\mathcal{R}_{min}$. The constraint $(C2)$ is to ensure that a device is associate only with one AP at a time. In constraints $(C3)$ and $(C4)$, $d_i \in \theta_z$ means that the statement is only valid if and only if the $i^{th}$ device is in the coverage of the $z$ AP, for $z \in \{b_j, a_h\}$. Constraints $(C5)$ and $(C6)$ are to ensure that an RB or an RC is occupied only by at most one device from the same AP's set of associated device. Constraints $(C7)$ and $(C8)$ draw the relationship between the resource allocation and the user association dynamics, where a device is only allowed to assume an RB at a BS or an RC at a DA if and only if it is associated with that particular AP. Constraint $(C9)$ ensures that any device is not allowed to transmit at a power level higher than $Q_{max}$. The second-stage sub-problem $(C10)$ means that every served device should be associated with the right AP with the minimum total transmit power when the system utility in the first-stage sub-problem is maximized, which is reinforced by the constraint given in $(C11)$. Notice that, the second-stage sub-problem is also subject to the same constraints defined in $(C1)$ through $(C9)$, given by constraint $(C12)$.

**P1-1:**

$$\mathcal{U}^* = \max_{\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{S},\mathbf{Q}_b,\mathbf{Q}_a} \quad \left( \sum_{b_j \in \Phi_b} \sum_{d_i \in \Upsilon_{b_j}} x_{d_i,b_j} + \sum_{a_h \in \Phi_a} \sum_{d_i \in \Upsilon_{a_h}} y_{d_i,a_h} \right)$$

$$\text{s.t.} \quad x_{d_i,b_j} \sum_{k=1}^{K} s^k_{d_i,b_j} \mathcal{R}^k_{d_i,b_j} + y_{d_i,a_h} \mathcal{R}_{d_i,2} \geq \mathcal{R}_{min} \tag{C1}$$

$$\sum_{b_j \in \Phi_b} x_{d_i,b_j} + \sum_{a_h \in \Phi_a} y_{d_i,a_h} \leq 1 \quad \forall d_i \in \Phi_d \tag{C2}$$

$$\begin{cases} x_{d_i,b_j} \in \{0,1\}, & \text{if } d_i \in \theta_{b_j} \ \forall b_j \in \Phi_b \\ x_{d_i,b_j} = 0, & \text{otherwise} \end{cases} \tag{C3}$$

$$\begin{cases} y_{d_i,a_h} \in \{0,1\}, & \text{if } d_i \in \theta_{a_h} \ \forall a_h \in \Phi_a \\ y_{d_i,a_h} = 0, & \text{otherwise} \end{cases} \tag{C4}$$

$$\sum_{d_i \in \Upsilon_{b_j}} s^k_{d_i,b_j} \leq 1, \quad \forall b_j, \& k \tag{C5}$$

$$\sum_{d_i \in \Upsilon_{a_h}} v^n_{d_i,a_h} \leq 1, \quad \forall a_h, \& n \tag{C6}$$

$$s^k_{d_i,b_j} = [0, x_{d_i,b_j}], \quad \forall d_i, b_j, \& k \tag{C7}$$

$$v^n_{d_i,a_h} = [0, y_{d_i,a_h}], \quad \forall d_i, a_h, \& n \tag{C8}$$

$$0 \leq x_{d_i,b_j} \sum_{k=1}^{K} s^k_{d_i,b_j} q^k_{d_i,b_j} + y_{d_i,a_h} \sum_{n=1}^{N} v^n_{d_i,a_h} q^n_{d_i,a_h} \leq Q_{max}, \tag{C9}$$

**P1-2:**

$$\min_{\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{S},\mathbf{Q}_b,\mathbf{Q}_a} \quad \sum_{b_j \in \Phi_b} \sum_{d_i \in \Upsilon_{b_j}} x_{d_i,b_j} \sum_{k=1}^{K} s^k_{d_i,b_j} q^k_{d_i,b_j} + \sum_{a_h \in \Phi_a} \sum_{d_i \in \Upsilon_{a_h}} y_{d_i,a_h} \sum_{n=1}^{N} v^n_{d_i,a_h} q^n_{d_i,a_h} \tag{C10}$$

$$\max_{\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{S},\mathbf{Q}_b,\mathbf{Q}_a} \left( \sum_{b_j \in \Phi_b} \sum_{d_i \in \Upsilon_{b_j}} x_{d_i,b_j} + \sum_{a_h \in \Phi_a} \sum_{d_i \in \Upsilon_{a_h}} y_{d_i,a_h} \right) = \mathcal{U}^* \tag{C11}$$

$$(C1) - (C9) \tag{C12}$$

Notice that problem **P1** is a complex multi-objective mixed integer non-convex optimization problem as it contains discrete indicative binary variables as well as a continuous variable representing the transmission power of the devices in the network and has a non-linear constraint presented in the logarithmic form of the achievable data rate. Furthermore, as the optimization objectives are part of the constraints of each other, this creates a duality that adds another dimension of complexity to the problem, forbidding the use of simple convex optimization arguments to solve **P1**. One way of simplifying the problem is to reformulate it as a single objective optimization problem which can be done via the method of weighted sum [86], given as show in

problem **P2**.

**P2:**

$$\min_{\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{S},\mathbf{Q}_b,\mathbf{Q}_a} \quad \epsilon\left(\sum_{b_j\in\Phi_b}\sum_{d_i\in\Upsilon_{b_j}} x_{d_i,b_j} q_{d_i}^{b_j} + \sum_{a_h\in\Phi_a}\sum_{d_i\in\Upsilon_{a_h}} y_{d_i,a_h} q_{d_i}^{a_h}\right)$$
$$-(1-\epsilon)\left(\sum_{b_j\in\Phi_b}\sum_{d_i\in\Upsilon_{b_j}} x_{d_i,b_j} + \sum_{a_h\in\Phi_a}\sum_{d_i\in\Upsilon_{a_h}} y_{d_i,a_h}\right)$$

s.t.
$$x_{d_i,b_j}\sum_{k=1}^{K} s_{d_i,b_j}^{k}\mathcal{R}_{d_i,b_j}^{k} + y_{d_i,a_h}\mathcal{R}_{d_i,2} \geq \mathcal{R}_{min} \qquad (C1)$$

$$\sum_{b_j\in\Phi_b} x_{d_i,b_j} + \sum_{a_h\in\Phi_a} y_{d_i,a_h} \leq 1 \qquad (C2)$$

$$\begin{cases} x_{d_i,b_j}\in\{0,1\}, & \text{if } d_i\in\theta_{b_j}\ \forall b_j\in\Phi_b \\ x_{d_i,b_j}=0, & \text{otherwise} \end{cases} \qquad (C3)$$

$$\begin{cases} y_{d_i,a_h}\in\{0,1\}, & \text{if } d_i\in\theta_{a_h}\ \forall a_h\in\Phi_a \\ y_{d_i,a_h}=0, & \text{otherwise} \end{cases} \qquad (C4)$$

$$\sum_{d_i\in\Upsilon_{b_j}} s_{d_i,b_j}^{k} \leq 1, \qquad \forall b_j, \&k \qquad (C5)$$

$$\sum_{d_i\in\Upsilon_{a_h}} v_{d_i,a_h}^{n} \leq 1, \qquad \forall a_h, \&n \qquad (C6)$$

$$s_{d_i,b_j}^{k} = [0, x_{d_i,b_j}], \qquad \forall d_i, b_j, \&k \qquad (C7)$$

$$v_{d_i,a_h}^{n} = [0, y_{d_i,a_h}], \qquad \forall d_i, a_h, \&n \qquad (C8)$$

$$0 \leq x_{d_i,b_j} q_{d_i}^{b_j} + y_{d_i,a_h} q_{d_i}^{a_h} \leq Q_{max}, \qquad (C9)$$

In problem **P2**, $q_{d_i}^{b_j} = \sum_{k=1}^{K} s_{d_i,b_j}^{k} q_{d_i,b_j}^{k}$ is the total transmission power of the $i^{th}$ device from the $j^{th}$ BS, $q_{d_i}^{a} = \sum_{n=1}^{N} v_{d_i,a_h}^{n} q_{d_i,a_h}^{n}$ is the total transmission power of the $i^{th}$ device from the $h^{th}$ DA, and $\epsilon$ is a constant that satisfies the following inequality for problem **P2** be equivalent to problem **P1** (proof in Appendix C.2)

$$0 \leq \epsilon \leq \frac{1}{1 + \sum_{q_{d_i}\in\Phi_d} Q_{max}}. \qquad (5.2)$$

Although problem **P2** is a single objective optimization problem and is simpler than yet equivalent to problem **P1**, it is still quite complex and combinatorial in nature due to the nested dependency between the objectives and constraints. Furthermore, problem **P2** can be classified as mixed-integer non-convex optimization problem due to the multiplication between discrete and continuous variables. Thus, traditional convex approaches cannot be applied to it. In what

follows, we propose a novel approach to solve problem **P2**, and hence solve the problem based on the theory of coalition formation games, where the problem is divided into multiple simpler sub-problems that can be tackled one at a time.

## 5.3 User association, resource allocation, and power control algorithm

In this section, we present a novel **P**ower control, resource **A**llocation, **U**ser association, QoS **S**atisfaction and **E**nergy consumption optimization algorithm, refereed to as *PAUSE*, to solve problem **P2**. The key idea in *PAUSE* is to divide problem **P2** into multiple simpler sub-problems by means of Coalition formation game theory and then devise the appropriate optimization techniques to solve the resultant sub-problems. We start with some preliminaries on coalition formation games to highlight their suitability in the context of this work. We then present the *PAUSE* algorithm in a systematic manner.

### 5.3.1 Preliminaries on Coalition formation games

In coalition formation games, players attempt to cluster such that the benefit each player gains from the cooperative clustering decisions is maximized, compared to the gains when they selfishly maximize their individual utilities. A coalition formation game is defined by means of three attributes, namely: i) the set of all players (i.e., the set of devices, $\mathcal{D}$, in the network) participating in the game to form cooperative clusters; ii) the partition of these players into clusters, denoted by $\mathcal{P} = \{C_1, ..., C_T\}$, which is a collection of $T$ coalitions; and iii) the coalition value, $\mathcal{V}(C_j, \mathcal{P})$, which quantifies the gain of the $j^{th}$ coalition, $C_j$, given the partition $\mathcal{P}$. Notice that, a coalition formation game may allow coalitions to overlap, allowing players to be part of one or more coalitions at the same time. However, in this work, we only consider non-overlapping games such that $C_i \cap C_j = \emptyset$ for all $i \neq j$, and $\cup_{i=1}^{T} C_i = \mathcal{D}$.

The objective of a coalition formation game is to find a partition that maximizes the gain value of all of its coalitions. In other words, let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two possible partitions of the devices in our considered DA infused network. In this context, a partition is essentially equivalent to determining which devices associate with which AP. Accordingly, to determine if the set of device associations corresponding to partition $\mathcal{P}_1$ is better than the set of device associations corresponding to partition $\mathcal{P}_2$, the following condition must be met:

$$\sum_{j \in \{\Phi_b, \Phi_a\}} \mathcal{V}(C_j, \mathcal{P}_1) < \sum_{j \in \{\Phi_b, \Phi_a\}} \mathcal{V}(C_j, \mathcal{P}_2) \tag{5.3}$$

where the "less than" inequality is because, in the context of problem **P2**, the objective is to find the partition that minimizes the objective function.

### 5.3.2 Coalition formation game for problem P2

As per the definition of coalition formation games, we can see that it can be applied in the context of problem **P2**, which is essentially a clustering (coalition formation) problem, the objective of which is to find the best user association decisions such that a user is associated with the AP which satisfies its data rate requirement while minimizing its energy consumption. As we have two tiers of APs, namely BSs and DAs, we create a super set of access points $\mathcal{C} = \{c_1, c_2, ..., c_{\mathcal{B}}, c_{\mathcal{B}+1}, ..., c_{\mathcal{B}+\mathcal{A}}\}$, such that $c_j$ denotes the $j^{th}$ BS in the network, for $0 < j \leq \mathcal{B}$, and the $j^{th}$ DA in the network, for $\mathcal{B} < j \leq \mathcal{B} + \mathcal{A}$. For ease of notation, let $\mathcal{J} = \mathcal{B} + \mathcal{A}$. Accordingly, let $\mathcal{P}_i = \{C_{1,i}, C_{2,i}, ..., C_{\mathcal{J},i}, C_{\mathcal{J}+1,i}\}$ denote the $i^{th}$ possible partition of the set of devices, $\mathcal{D}$, in the network, where $C_{j,i}$ denotes the set of devices associated with the $j^{th}$ AP, and $C_{\mathcal{J}+1,i}$ is an extra coalition in which devices that are not yet associated with any AP are placed. Since, in problem **P2**, constraint (C2) enforces that a devices is associated with only one AP at a time, we have a non-overlapping coalition formation game where $C_{j,i} \cap C_{j',i} = \emptyset$ for all $j \neq j'$, and $\sum_{j=1}^{\mathcal{J}+1} C_{j,i} = \mathcal{D}$. It should be noted that the coalition $C_{\mathcal{J}+1}$ can also be thought of as the set of devices that cannot be served at the required data rate requirement given the limitation on the maximum transmission power of the devices as well as the limited resources at the APs.

As per the preceding model, we can now define the gain value function for problem **P2**, $\mathcal{V}(C_{j,i}, \mathcal{P}_i)$, for the $j^{th}$ coalition $C_{j,i}$ given the $i^{th}$ partition $\mathcal{P}_i$, to be given as

$$\mathcal{V}(C_{j,i}, \mathcal{P}_i) = \begin{cases} \epsilon \sum\limits_{d^i_{k,j} \in C_{j,i}} \widetilde{q}_{d^i_{k,j}} - (1-\epsilon)|C_{j,i}|, & \text{if } 0 < j \leq \mathcal{J} \\ 0, & \text{if } j = \mathcal{J}+1 \end{cases} \tag{5.4}$$

where $\widetilde{q}_{d^i_{k,j}}$ denotes the optimal transmission power of the $k^{th}$ device from the $j^{th}$ AP given the $i^{th}$ partition, and $|C_{j,i}|$ is the cardinality of the $j^{th}$ coalition from the $i^{th}$ partition (i.e., the number of satisfied devices associated with the $j^{th}$ AP given the partition $\mathcal{P}_i$).

### 5.3.3 Resource allocation and transmission power

We can deduce that the outcome of the coalition formation game is a set of coalitions, $\{C_1, ..., C_{\mathcal{J}+1}\}$, which determines the association decisions of the devices in the network. Accordingly, given partition $\mathcal{P} = \{C_1, C_2, ..., C_{\mathcal{J}+1}\}$, we can say that $x_{d_i,b_j} = 1$ if $d_i \in C_j$ and zero otherwise, for $0 < j \leq \mathcal{B}$, and $y_{d_i,a_h} = 1$ if $d_i \in C_j$ and zero otherwise, for $\mathcal{B} < j \leq \mathcal{J}$. Also, both $x_{d_i,b_j}$ and $y_{d_i,a_h}$ equal to zero if $d_i \in C_{\mathcal{J}+1}$. In other words, the coalition formation game produces

user association matrices $X$ and $Y$ for different possible partitions. Once we have the association matrices corresponding to a particular partition, their optimality shall be tested against the objective function of problem **P2**. To do so, we need to prove that the partition, resulting in those matrices, minimizes the value gain function defined in (5.4), for all of its coalitions. As the first part of the value function is the summation of the transmission power of all devices in a coalition, for the given partition, we need to find the corresponding optimal resource allocations and minimum power assignment decisions that satisfy the data rate requirements of the devices. Mathematically, given the partition $\mathcal{P}_h$, the optimal transmission power of the devices and RBs and RCs assignments can be written as an minimization problem, given by

**P3:**

$$\min_{\mathbf{V},\mathbf{S},\mathbf{Q}_b,\mathbf{Q}_a} \quad \sum_{d_i \in C_{1,h}}^{C_{\mathcal{B},h}} \sum_{k=1}^{K} s^k_{d_i,C_{j,h}} q^k_{d_i,C_{j,h}} + \sum_{d_i \in C_{\mathcal{B}+1}}^{C_{\mathcal{J}}} \sum_{n=1}^{N} v^n_{d_i,C_j} q^n_{d_i,C_{j,h}}$$

$$\text{s.t.} \quad \begin{cases} \sum_{k=1}^{K} s^k_{d_i,C_{j,h}} \mathcal{R}^k_{d_i,C_{j,h}} \geq \mathcal{R}_{min}, & \text{if } j \leq \mathcal{B} \\ \min\left(\mathcal{R}_{a_j}, \sum_{n=1}^{N} v^n_{d_i,C_{j,h}} \mathcal{R}^n_{d_i,C_{j,h}}\right) \geq \mathcal{R}_{min}, & \text{if } \mathcal{B} < j \leq \mathcal{J} \\ 0, & \text{if } j = \mathcal{J}+1 \end{cases} \qquad (C1)$$

$$\sum_{d_i \in C_{j,h}} s^k_{d_i,C_{j,h}} \leq 1, \qquad \text{for } j \neq \mathcal{J}+1 \qquad (C2)$$

$$\sum_{d_i \in C_{j,h}} v^n_{d_i,C_{j,h}} \leq 1, \qquad \text{for } j \neq \mathcal{J}+1 \qquad (C3)$$

$$s^k_{d_i,C_{j,h}} = \begin{cases} [0,1], & \text{for } j \leq \mathcal{B} \\ 0, & \text{otherwise} \end{cases} \qquad (C4)$$

$$v^n_{d_i,C_{j,h}} = \begin{cases} [0,1], & \text{for } \mathcal{B} < j \leq \mathcal{J} \\ 0, & \text{otherwise} \end{cases} \qquad (C5)$$

$$0 \leq \sum_{k=1}^{K} s^k_{d_i,C_{j,h}} q^k_{d_i,C_{j,h}} \leq Q_{max}, \qquad \forall d_i \in \Phi_d \text{ for } j < \mathcal{B} \qquad (C6)$$

$$0 \leq \sum_{n=1}^{N} v^n_{d_i,C_{j,h}} q^n_{d_i,C_{j,h}} \leq Q_{max}, \qquad \forall d_i \in \Phi_d \text{ for } \mathcal{B} < j \leq \mathcal{J} \qquad (C7)$$

where

$$R^k_{d_i,C_{j,h}} = \omega_b \log_2\left(1 + \frac{q^k_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^k_{d_{i(j)},C_{j,h}}}{\sum_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum_{d_n \in C_{m,h}} s^k_{d_n,C_{m,h}} q^k_{d_n,C_{m,h}} ||d_n - c_m||^{-\alpha} g^k_{d_{n(m)},b_j}}\right),$$

$$\text{for } 0 < \ j \ \& \ m \leq \mathcal{B}. \qquad (5.5)$$

$$R^n_{d_i,C_{j,h}} = \omega_c \log_2\left(1 + \frac{q^n_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^n_{d_{i(j)},C_{j,h}}}{\sum\limits_{C_{m,h}\in\mathcal{P}_h/C_{j,h}} \sum\limits_{d_l\in C_{m,h}} v^n_{d_l,C_{m,h}} q^n_{d_l,C_{m,h}} ||d_l - c_m||^{-\alpha} g^n_{d_{l(m)},a_j}}\right),$$

$$\text{for } \mathcal{B} < j \;\&\; m \leq \mathcal{J}. \tag{5.6}$$

As we consider that channels $C^1_d$ and $C^2_d$ are orthogonal, transmissions from devices to the DAs and those directly to the BSs do not interfere. Accordingly, there is no dependency between the RB and RC assignment or the uplink transmission power of directly connected devices and those transmitting towards DAs. This independence permits the division of problem **P3** into two independent yet similar optimization problems, one for BS-directly connected devices and the other for devices connected to DAs. Those optimization problems can be written as (given the partition $\mathcal{P}_h$)

**P4-1:**
$$\min_{\mathbf{S},\mathbf{Q}_b} \quad \sum_{d_i\in C_{1,h}}^{C_{\mathcal{B},h}} \sum_{k=1}^{K} s^k_{d_i,C_{j,h}} q^k_{d_i,C_{j,h}}$$

s.t.
$$\sum_{k=1}^{K} s^k_{d_i,C_{j,h}} \mathcal{R}^k_{d_i,C_{j,h}} \geq \mathcal{R}_{min}, \qquad \text{for } j \leq \mathcal{B} \qquad (C4-1.1)$$

$$\sum_{d_i\in C_{j,h}} s^k_{d_i,C_{j,h}} \leq 1, \qquad \text{for } j \neq \mathcal{J}+1 \qquad (C4-1.2)$$

$$s^k_{d_i,C_{j,h}} = [0,1], \qquad \text{for } j \leq \mathcal{B} \qquad (C4-1.3)$$

$$0 \leq \sum_{k=1}^{K} s^k_{d_i,C_{j,h}} q^k_{d_i,C_{j,h}} \leq Q_{max}, \qquad \text{for } j < \mathcal{B} \qquad (C4-1.4)$$

**P4-2:**
$$\min_{\mathbf{V},\mathbf{Q}_a} \quad \sum_{d_i\in C_{\mathcal{B}+1}}^{C_{\mathcal{J}}} \sum_{n=1}^{N} v^n_{d_i,C_j} q^n_{d_i,C_{j,h}}$$

s.t.
$$\sum_{n=1}^{N} v^n_{d_i,C_{j,h}} \mathcal{R}^n_{d_i,C_{j,h}} \geq \mathcal{R}_{min}, \quad \text{if } \mathcal{R}_{a_j} > \mathcal{R}_{min}, \quad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C4-2.1)$$

$$\sum_{d_i\in C_{j,h}} v^n_{d_i,C_{j,h}} \leq 1, \qquad \text{for } j \neq \mathcal{J}+1 \qquad (C4-2.2)$$

$$v^n_{d_i,C_{j,h}} = [0,1], \qquad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C4-2.3)$$

$$0 \leq \sum_{n=1}^{N} v^n_{d_i,C_{j,h}} q^n_{d_i,C_{j,h}} \leq Q_{max}, \qquad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C4-2.4)$$

It should be noted that different from problem **P4-1**, problem **P4-2** is only valid if and only if a device is associated with a DA that achieves a data rate of at least $\mathcal{R}_{min}$ as indicated by constraint $(C4-2.1)$.

To be able to solve problem **P3**, the partition $\mathcal{P}_h = \{C_{1,h}, C_{2,h}, ..., C_{\mathcal{J},h}, C_{\mathcal{J}+1,h}\}$ should be feasible. In other words, given the association matrices $\mathbf{X}$ and $\mathbf{Y}$ corresponding to partition $\mathcal{P}_h$,

both problems **P4-1** and **P4-2** should have a possible solution. That is, given $\mathbf{X}$, there is a set of possible RB assignments, $\mathbf{S}$, and transmission power control levels, $\mathbf{Q}_b$, that satisfies the data rate requirements of each device in every coalition in partition $\mathcal{P}_h$ for $0 < j \leq \mathcal{B}$. Similarly, given $\mathbf{Y}$, there is a set of possible RC assignments, $\mathbf{V}$, and transmission power control levels, $\mathbf{Q}_a$, that satisfies the data rate requirements of each device in every coalition in partition $\mathcal{P}_h$ for $\mathcal{B} < j \leq \mathcal{J}$. Otherwise, partition $\mathcal{P}_h$ is not feasible and other feasible partitions should be found.

Notice that both problems **P4-1** and **P4-2** are typical joint power minimization and resource allocation problems that have been tackled in the literature quite often. As the literature suggests, these two problems can be solved optimally using Branch and Bound, and outer Approximation methods [72]. In fact, one popular way of solving this problem is using generalized Bender's decomposition (GBD) [94]. However, due to the non-convex nature of the problem and the fact that it is an MINLP, solving this problem optimally might not be feasible in polynomial time. Thus, in what follows, we propose a sub-optimal resource allocation and power minimization algorithm to solve problems **P4-1** and **P4-2**. We start by reformulating the problem as a difference of two concave functions programming (D.C. programming). Although there is no analytic evidence to justify this, many results from the literature show that with the proper choice of the starting point, D.C. programming algorithms can find a local optimal point that often yields the global optimum. A number of regularization and starting-point choosing methods have been proposed and can be used in this work to ensure finding the global optimum. Yet, our initial objective from solving problems **P4-1** and **P4-2** is to prove the feasibility of the given partitions from problem **P3**. The second step is to find the corresponding optimal power assignment and resource allocations once the *PAUSE* algorithm converges to a feasible partition.

### 5.3.4 Sub-optimal reformulation as D.C. programming

Both problems **P4-1** and **P4-2** are highly non-convex due to constraints ($C4-1.1$) and ($C4-2.1$). They are also classified as mixed integer non-linear programming problems due to the presence of both continuous and discrete variables. As both $s^k_{d_i,C_{j,h}}$ and $v^n_{d_i,C_{j,h}}$ are binary variables, its worth pointing out the following equalities

$$s^k_{d_i,C_{j,h}} R^k_{d_i,C_{j,h}} = \omega_b \log_2\left(1 + \frac{s^k_{d_i,C_{j,h}} q^k_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^k_{d_{i(j)},c_j}}{\sum\limits_{C_{m,h}\in\mathcal{P}_h/C_{j,h}} \sum\limits_{d_n\in C_{m,h}} s^k_{d_n,C_{m,h}} q^k_{d_n,C_{m,h}} ||d_n - c_m||^{-\alpha} g^k_{d_{n(m)},c_j}}\right), \quad \text{for } 0 < \ j \ \& \ m \leq \mathcal{B}. \quad (5.7)$$

$$v^n_{d_i,C_{j,h}} R^n_{d_i,C_{j,h}} = \omega_c \log_2\left(1 + \frac{v^n_{d_i,C_{j,h}} q^n_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^n_{d_{i(j)},c_j}}{\sum\limits_{C_{m,h}\in\mathcal{P}_h/C_{j,h}} \sum\limits_{d_l\in C_{m,h}} v^n_{d_l,C_{m,h}} q^n_{d_l,C_{m,h}} ||d_l - c_m||^{-\alpha} g^n_{d_{l(m)},c_j}}\right),$$

$$\text{for } \mathcal{B} < j \;\&\; m \leq \mathcal{J}. \quad (5.8)$$

One way of simplifying problems **P4-1** and **P4-2** is to redefine constraints $(C4-1.3)$ and $(C4-2.3)$ with an equivalent continuous representation while enforcing that variables $s^k_{d_i,C_{j,h}}$ and $v^n_{d_i,C_{j,h}}$ assume binary values. To do so, first, recognize that constraints $(C4-1.3)$ and $(C4-2.3)$ can be expressed as the intersection of the following regions:

$$(C4-1.3) \Rightarrow \begin{cases} R^1_{4-1} : 0 \leq s^k_{d_i,C_{j,h}} \leq 1, & \forall i,j,k, \quad (C4-1.3.1) \\ R^2_{4-1} : \sum\limits_j \sum\limits_i \sum\limits_k \left(s^k_{d_i,C_{j,h}} - (s^k_{d_i,C_{j,h}})^2\right) \leq 0, & (C4-1.3.2) \end{cases} \quad (5.9)$$

$$(C4-2.3) \Rightarrow \begin{cases} R^1_{4-2} : 0 \leq v^n_{d_i,C_{j,h}} \leq 1, & \forall i,j,n, \quad (C4-2.3.1) \\ R^2_{4-2} : \sum\limits_j \sum\limits_i \sum\limits_n \left(v^n_{d_i,C_{j,h}} - (v^n_{d_i,C_{j,h}})^2\right) \leq 0. & (C4-2.3.2) \end{cases} \quad (5.10)$$

Second, as constraints $(C4-1.1)$ and $(C4-2.1)$ are mixed integers in nature, by reformulating constraints $(C4-1.4)$ and $(C4-1.4)$ such that their mixed integer nature is replaced by continuous constraints, and utilizing the redefinition of constraints $(C4-1.3)$ and $(C4-2.3)$ given in (5.9) and (5.10) respectively, and with the help of Lagrangian methodology [19, 63], problems **P4-1** and **P4-2** can be rewritten as problems **P5-1** and **P5-2** shown below.

**P5-1:**

$$\min_{\mathbf{S},\mathbf{Q}_b} \quad \sum_{d_i\in C_{1,h}}^{C_{\mathcal{B},h}} \sum_{k=1}^{K} q^k_{d_i,C_{j,h}} + \mu_1 \sum_j \sum_i \sum_k \left(s^k_{d_i,C_{j,h}} - (s^k_{d_i,C_{j,h}})^2\right)$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \widetilde{\mathcal{R}}^k_{d_i,C_{j,h}} \geq \mathcal{R}_{min}, \quad \text{for } j \leq \mathcal{B} \quad (C5-1.1)$$

$$\sum_{d_i\in C_{j,h}} s^k_{d_i,C_{j,h}} \leq 1, \quad \text{for } j \neq \mathcal{J}+1 \quad (C5-1.2)$$

$$0 \leq s^k_{d_i,C_{j,h}} \leq 1, \quad \text{for } j \leq \mathcal{B} \quad (C5-1.3)$$

$$0 \leq \sum_{k=1}^{K} q^k_{d_i,C_{j,h}} \leq Q_{max}, \quad \text{for } j < \mathcal{B} \quad (C5-1.4)$$

$$0 \leq q^k_{d_i,C_{j,h}} \leq s^k_{d_i,C_{j,h}} Q_{max}, \quad \text{for } j \leq \mathcal{B} \quad (C5-1.5)$$

**P5-2:**

$$\min_{\mathbf{v},\mathcal{P}_h,\mathbf{Q}_a} \quad \sum_{d_i \in C_{\mathcal{B}+1}}^{C_{\mathcal{J}}} \sum_{n=1}^{N} q^n_{d_i,C_{j,h}} + \mu_2 \sum_j \sum_i \sum_n \left( v^n_{d_i,C_{j,h}} - (v^n_{d_i,C_{j,h}})^2 \right)$$

$$\text{s.t.} \quad \sum_{n=1}^{N} \widetilde{\mathcal{R}}^n_{d_i,C_{j,h}} \geq \mathcal{R}_{min}, \qquad \text{if } \mathcal{R}_{a_j} > \mathcal{R}_{min}, \text{ for } \mathcal{B} < j \leq \mathcal{J} \qquad (C5-2.1)$$

$$\sum_{d_i \in C_{j,h}} v^n_{d_i,C_{j,h}} \leq 1, \qquad \text{for } j \neq \mathcal{J}+1 \qquad (C5-2.2)$$

$$0 \leq v^n_{d_i,C_{j,h}} \leq 1, \qquad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C5-2.3)$$

$$0 \leq \sum_{n=1}^{N} q^n_{d_i,C_{j,h}} \leq Q_{max}, \qquad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C5-2.4)$$

$$0 \leq q^n_{d_i,C_{j,h}} \leq v^n_{d_i,C_{j,h}} Q_{max}, \qquad \text{for } \mathcal{B} < j \leq \mathcal{J} \qquad (C5-1.5)$$

In problems **P5-1** and **P5-2**, $\mu_1 >> 1$ and $\mu_2 >> 1$ are Lagrangian multipliers which define the penalties when variables $s^k_{d_i,C_{j,h}}$ and $v^n_{d_i,C_{j,h}}$ are set to values other than 0 or 1, ensuring that problems **P5-1** and **P5-2** are equivalent to problems **P4-1** and **P4-2** respectively.

For compactness, in what follows, we use $\mathcal{W}_1$ and $\mathcal{W}_2$ to denote the set of constraints $C5-1.2$ - $C5-1.5$ and $C5-2.2$ - $C5-2.5$ respectively. Further, for simplicity of notation, let $\Omega_1(\mathbf{Q}_b) = \sum_{d_i \in C_{1,h}}^{C_{\mathcal{B},h}} \sum_{k=1}^{K} q^k_{d_i,C_{j,h}}$ and $\Omega_2(\mathbf{Q}_a) = \sum_{d_i \in C_{\mathcal{B}+1,h}}^{C_{\mathcal{J},h}} \sum_{n=1}^{N} q^N_{d_i,C_{j,h}}$. Moreover, $\widetilde{\mathcal{R}}^k_{d_i,C_{j,h}}$ and $\widetilde{\mathcal{R}}^n_{d_i,C_{j,h}}$ are given as

$$\widetilde{\mathcal{R}}^k_{d_i,C_{j,h}} = \omega_b \log_2 \left( 1 + \frac{q^k_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^k_{d_{i(j)},c_j}}{\sum\limits_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum\limits_{d_n \in C_{m,h}} q^k_{d_n,C_{m,h}} ||d_n - c_m||^{-\alpha} g^k_{d_{n(m)},c_j}} \right), \qquad \text{for } 0 < \ j \ \& \ m \leq \mathcal{B}.$$

$$\widetilde{\mathcal{R}}^n_{d_i,C_{j,h}} = \omega_c \log_2 \left( 1 + \frac{q^n_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^n_{d_{i(j)},c_j}}{\sum\limits_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum\limits_{d_l \in C_{m,h}} q^n_{d_l,C_{m,h}} ||d_l - c_m||^{-\alpha} g^n_{d_{l(m)},c_j}} \right), \qquad \text{for } \mathcal{B} < j \ \& \ m \leq \mathcal{J}.$$

*Proposition 1: For sufficiently constant positive large values of $\mu_1$ and $\mu_2$, problems* **P4-1** *and* **P4-2** *are respectively equivalent to problems* **P5-1** *and* **P5-2**.

*Proof.* Please refer to Appendix C.3 for detailed proof of Proposition 1.

Last, notice that, both problems **P5-1** and **P5-2** are now functions of only continuous variables, yet constraints $(C4-1.1)$ and $(C4-2.1)$ are highly non-convex. These non-convex constraints can be reformulated into a difference between two concave functions [63, 98], given as

$$\sum_{k=1}^{K} \widetilde{\mathcal{R}}^{k}_{d_i,C_{j,h}} =$$

$$\underbrace{\omega_b \sum_{k=1}^{K} \log_2 \left( q^{k}_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^{k}_{d_{i(j)},c_j} + \sum_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum_{d_n \in C_{m,h}} q^{k}_{d_n,C_{m,h}} ||d_n - c_m||^{-\alpha} g^{k}_{d_{n(m)},c_j} \right)}_{\eta_b(\mathbf{Q}_b)}$$

$$\underbrace{-\omega_b \sum_{k=1}^{K} \log_2 \left( \sum_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum_{d_n \in C_{m,h}} q^{k}_{d_n,C_{m,h}} ||d_n - c_m||^{-\alpha} g^{k}_{d_{n(m)},c_j} \right)}_{\zeta_b(\mathbf{Q}_b)}. \tag{5.11}$$

$$\sum_{n=1}^{N} \widetilde{\mathcal{R}}^{n}_{d_i,C_{j,h}} =$$

$$\underbrace{\omega_c \sum_{n=1}^{N} \log_2 \left( q^{n}_{d_i,C_{j,h}} ||d_i - c_j||^{-\alpha} g^{n}_{d_{i(j)},c_j} + \sum_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum_{d_l \in C_{m,h}} q^{n}_{d_l,C_{m,h}} ||d_l - c_m||^{-\alpha} g^{n}_{d_{l(m)},c_j} \right)}_{\eta_a(\mathbf{Q}_a)}$$

$$\underbrace{-\omega_b \log_2 \left( \sum_{C_{m,h} \in \mathcal{P}_h / C_{j,h}} \sum_{d_l \in C_{m,h}} q^{n}_{d_l,C_{m,h}} ||d_l - c_m||^{-\alpha} g^{n}_{d_{l(m)},c_j} \right)}_{\zeta_a(\mathbf{Q}_a)}. \tag{5.12}$$

Thus, problems **P5-1** and **P5-2** can be written as

**P6-1:**

$$\min_{\mathbf{S},\mathcal{P}_h,\mathbf{Q}_b} \Omega_1(\mathbf{Q}_b) + \mu_1 \sum_j \sum_i \sum_k \left( s^{k}_{d_i,C_{j,h}} - (s^{k}_{d_i,C_{j,h}})^2 \right)$$

$$\text{s.t.} \quad \mathcal{W}_1, \ \ \omega_b(\eta_b(\mathbf{Q}_b) - \zeta_b(\mathbf{Q}_b)) \geq \mathcal{R}_{min}, \qquad \text{for } j \leq \mathcal{B} \qquad (C6-1.1)$$

**P6-2:**

$$\min_{\mathbf{V},\mathcal{P}_h,\mathbf{Q}_a} \Omega_2(\mathbf{Q}_a) + \mu_2 \sum_j \sum_i \sum_n \left( v^{n}_{d_i,C_{j,h}} - (v^{n}_{d_i,C_{j,h}})^2 \right)$$

$$\text{s.t.} \quad \mathcal{W}_2, \ \ \omega_c(\eta_a(\mathbf{Q}_a) - \zeta_a(\mathbf{Q}_a)) \geq \mathcal{R}_{min}, \ \ \text{if } \mathcal{R}_{a_j} > \mathcal{R}_{min}, \text{ for } \mathcal{B} < j \leq \mathcal{J} \qquad (C6-2.1)$$

Problems **P6-1** and **P6-2** are in the form of the difference of two concave (D.C) functions. They can be solved using iterative approaches, starting from an initial point and employing the first order Taylor approximation to change the concave problems to convex ones. Note that, as problems **P6-1** and **P6-2** are in their canonical form of D.C. programming [19], in particular, terms $u(\mathbf{S}) = \sum_j \sum_i \sum_k (s^{k}_{d_i,C_{j,h}})^2$, $u(\mathbf{V}) = \sum_j \sum_i \sum_n (v^{n}_{d_i,C_{j,h}})^2$, $\zeta_b(\mathbf{Q}_b)$ and $\zeta_a(\mathbf{Q}_a)$ are all concave functions and the rest of the constraints are convex, techniques such as successive

convex approximation can be used to arrive at stationary points for their solutions [16].

### 5.3.5 Iterative algorithm for the sub-optimal formulation

The following analysis is applicable for both problems **P6-1** and **P6-2**. However, for brevity, we focus only on **P6-1**.

To be able to solve problem **P6-1**, the concavity introduced by the Logarithm and square functions can be transformed into a equivalent convex form by recognizing that both the terms $u(\mathbf{S})$ and $\zeta_b(\mathbf{Q}_b)$ are differentiable. Thus, using their first order Taylor expansion, the following global underestimators, shown by the inequalities, hold for any feasible solution point $(\mathbf{Q}_b^t, \mathbf{S}^t)$.

$$u(\mathbf{S}) \quad \geq u(\mathbf{S}^t) + \nabla_{\mathbf{S}} u(\mathbf{S}^t)(\mathbf{S} - \mathbf{S}^t) \tag{5.13}$$

$$\zeta_b(\mathbf{Q}_b) \geq \zeta_b(\mathbf{Q}_b^t) + \nabla_{\mathbf{Q}_b} \zeta_b(\mathbf{Q}_b^t)(\mathbf{Q}_b - \mathbf{Q}_b^t) \tag{5.14}$$

where $\nabla_X$ denotes the partial derivatives of the function with respect to the vector $X$, and $\mathbf{Q}_b^t$ and $\mathbf{S}^t$ are the feasible solutions of problem **P7-1** at the $t^{th}$ iteration of the iterative algorithm.

By taking the differentials of the Logarithmic constraints and the squared portion of the objective functions of problems **P7-1** and **P7-2**, we can redefine them in a convex manner, given as

**P7-1:**

$$\min_{\mathbf{S}, \mathcal{P}_h, \mathbf{Q}_b} \Omega_1(\mathbf{Q}_b) + \mu_1 \sum_j \sum_i \sum_k \left( s^k_{d_i, C_{j,h}} - ((s^k_{d_i, C_{j,h}})^t)^2 - 2(s^k_{d_i, C_{j,h}})^t ((s^k_{d_i, C_{j,h}}) - (s^k_{d_i, C_{j,h}})^t) \right)$$

$$\text{s.t.} \quad \mathcal{W}_1,\ \omega_b(\eta_b(\mathbf{Q}_b) - (\zeta_b(\mathbf{Q}_b^t) + \nabla_{\mathbf{Q}_b} \zeta_b(\mathbf{Q}_b^t)(\mathbf{Q}_b - \mathbf{Q}_b^t))) \geq \mathcal{R}_{min}, \qquad \text{for } j \leq \mathcal{B}$$

**P7-2:**

$$\min_{\mathbf{V}, \mathcal{P}_h, \mathbf{Q}_a} \Omega_2(\mathbf{Q}_a) + \mu_2 \sum_j \sum_i \sum_n \left( v^n_{d_i, C_{j,h}} - ((v^n_{d_i, C_{j,h}})^t)^2 - 2(v^n_{d_i, C_{j,h}})^t ((v^n_{d_i, C_{j,h}}) - (v^n_{d_i, C_{j,h}})^t) \right)$$

$$\text{s.t.} \quad \mathcal{W}_2,\ \omega_c(\eta_a(\mathbf{Q}_a) - (\zeta_a(\mathbf{Q}_a^t) + \nabla_{\mathbf{Q}_a} \zeta_a(\mathbf{Q}_a^t)(\mathbf{Q}_a - \mathbf{Q}_a^t))) \geq \mathcal{R}_{min},\ \text{ if } \mathcal{R}_{a_j} > \mathcal{R}_{min},\ \text{for } \mathcal{B} < j \leq \mathcal{J}$$

Our method to solve the redefined convex optimization problems **P7-1** and **P7-2** is shown in Algorithm 1. As both problems are the same, for brevity, Algorithm 1 is given in the notations of problem **P7-1**.

where $||\cdot||_1$ is the 1-norm of the argument which is the sum of the absolute values of the columns of the argument matrix (Manhattan norm).

As per the inequalities given in (5.13) and (5.14), problems **P7-1** and **P7-2** represent an upper bound to problems **P6-1** and **P6-2** respectively. The iterative method presented in Algorithm 1 tightens the upper bound depending on the stopping condition. In Algorithm 1, we

**Algorithm 1** Iterative D.C. programming algorithm for problems **P7-1** and **P7-2**

1: Initialize the stopping criteria tolerance $e > 0$, the maximum number of iterations $T_{max}$, $t = 0$, the penalty factor $\mu_1 >> 0$, and feasible matrices $\mathbf{Q}_b^0$ and $\mathbf{S}^0$
2: **Iteration** $t$: For a given point $\Gamma^t = (\mathbf{Q}_b^t, \mathbf{S}^t)$, execute the following
3: Step 1: Compute $\nabla_{\mathbf{Q}_b} \zeta_b(\mathbf{Q}_b^t)$ and $\nabla_{\mathbf{S}} u(\mathbf{S}^t)$
4: Step 2: Solve the convex problem **P7-1** to find the temporal solution $\widetilde{\Gamma} = (\widetilde{\mathbf{Q}}_b, \widetilde{\mathbf{S}})$
5: Step 3: if $||\widetilde{\mathbf{Q}}_b||_1 - ||\mathbf{Q}_b^t||_1 \leq e$, then stop solving problem **P7-1**. Set $\mathbf{Q} = \widetilde{\mathbf{Q}}_b$ and $\mathbf{S} = \widetilde{\mathbf{S}}$.
Otherwise, $\Gamma^{t+1} = \widetilde{\Gamma}$ and go back to step 1

start with an initial point $\Gamma^0$ which is used to solve the convex problems **P7-1** and **P7-2** for an updated temporal solution $\widetilde{\Gamma}$. The updated temporal solution is then recursively used to find another temporal solution. The update process continues until the algorithm converges. Algorithm 1 results in a series of solutions to problems **P7-1** and **P7-2** by solving their equivalent convex counterparts. The work presented in [100] can be used to show that Algorithm 1 converges locally to a stationary point with polynomial time computational complexity; however, for completeness, the proof of convergence is given in Appendix C.4.

## 5.4 The *PAUSE* algorithm

In what follows, we describe the working process of the proposed algorithm *PAUSE* to solve problem **P2** which consists of two main stages: the initial partition creation stage and the partition update stage. Below we give a detailed description of each stage as well as present the corresponding pseudo-codes making up the full *PAUSE* algorithm.

### 5.4.1 Algorithm description

#### 5.4.1.1 Initialization

The key objective of the *PAUSE* algorithm is to find an optimal partition of the devices that results in the maximization of the number of supported and QoS satisfied devices while optimizing resource utilization and minimizing the total transmission power of the devices. Consequently, in the first step of *PAUSE*, devices in the network are divided into coalitions each associated with an AP in the network. Thus, we first need to choose an initial feasible partition. Let the initial partition, at $h = 0$, be donated by $\mathcal{P}_0 = \{C_{1,0}, ..., C_{\mathcal{J}+1,0}\}$ where $C_{j,0}$ denotes the coalition associated with the $j^{th}$ BS, for $0 < j \leq \mathcal{B}$, and the $j^{th}$ DA, for $\mathcal{B} < j \leq \mathcal{J}$, and an extra coalition that belongs to no AP, index by $j = \mathcal{J} + 1$. Initially, all devices are assumed to be in the coalition $C_{\mathcal{J}+1,0}$; i.e., $C_{j,0} = \emptyset$ for $0 < j \leq \mathcal{J}$ and $C_{\mathcal{J}+1,0} = \mathcal{D}$. Sequentially, we take every device from $C_{\mathcal{J}+1,0}$ and randomly place it into one of the other $\mathcal{J}$ coalitions. For a device to be

placed in a coalition, it has to be in the coverage of the AP point belonging to that coalition (i.e., $d_i \in \theta_j$ for $j \in \{\Phi_b, \Phi_a\}$). Once a device is moved from $C_{\mathcal{J}+1,0}$, we update the partition $\mathcal{P}_0$ and perform a partition feasibility check as explained next.

#### 5.4.1.2 Feasibility check

In the initialization stage, it takes at least $\mathcal{D}$ movements to parse through all the devices in the network and place them into a feasible partition. For the $i^{th}$ device moved out of $C_{\mathcal{J}+1,0}$ and placed into a feasible coalition, $C_{j,0}$, an intermediate partition is formed. Let $\mathcal{P}_0^i$ denote the intermediate partition after the placement of the $i^{th}$ device into a feasible coalition. We derive the association matrices $\mathrm{X}_0^i$ and $\mathrm{Y}_0^i$ corresponding to $\mathcal{P}_0^i$ and pass them to Algorithm 1. If a solution is attainable, i.e., given $\mathrm{X}_0^i$ and $\mathrm{Y}_0^i$ problems **P7-1** and **P7-2** have solutions, then the corresponding placement of the $i^{th}$ device into the $j^{th}$ coalition is approved, for $j \leq \mathcal{J}$. If Algorithm 1 does not converge, the device movement from $C_{\mathcal{J}+1,0}$ into $C_{j,0}$ is not approved and the device stays in $C_{\mathcal{J}+1,0}$. The same process is repeated sequentially for all the devices in $C_{\mathcal{J}+1,0}$. After $\mathcal{D}$ steps, if $C_{\mathcal{J}+1,0} = \emptyset$, the feasibility check stage is terminated; otherwise, the placement procedure is re-execute for the devices that were returned to $C_{\mathcal{J}+1,0}$. Let $\vartheta$ be the number of times the re-execution of the feasibility check stage after the $\mathcal{D}^{th}$ time. To guarantee the termination of the feasibility check stage, we consider that the re-execution is done with probability $\rho = 1 - \exp\{-\kappa\vartheta\}$, where $0 < \kappa < 1$ is a control parameter for the decay speed of $\rho$. The process keeps iterating until the condition $\rho < \varrho$ is met, where $0 \leq \varrho \leq 1$ is a design threshold parameter for terminating the feasibility check stage. As the objective is to maximize the number of supported devices, we seek that the initial partition has as many devices as possible distributed across the APs. To do so, the threshold $\varrho$ can aggressively be set to a small value to ensure as many re-executions as possible, however, this comes at the expense of increased algorithm convergence time. The procedure of creating the initial feasible partition is summarized in Algorithm 2.

#### 5.4.1.3 Partition update

Once the feasibility check stage is terminated, the final initial feasible partition is generated, donated by $\mathcal{P}_0$, and the partition update stage commences. The update stage can be generalized as follows. At the $h^{th}$ iteration of the partition update, a mutation of the $\mathcal{P}_{h-1}$ partition, denoted by $\widetilde{\mathcal{P}}_{h-1}$, is created by randomly choosing a device from two different coalitions $C_{j,0}$ and $C_{j',0}$, for $j \neq j'$ and $0 < j \;\&\; j' \leq \mathcal{J}+1$, and swapping the devices together. One of the three following scenarios occurs: i) the chosen device from $C_{j,0}$ is not in the coverage of $C_{j',0}$ and that from

$C_{j',0}$ is not in the coverage of $C_{j,0}$, ii) one of the chosen devices can be swapped while the other cannot due to coverage constraints, or iii) both devices can be swapped as they are in coverage. For scenario one, the swap is rejected and the mutation is not approved; thus, $\mathcal{P}_h = \mathcal{P}_{h-1}$. For scenario two, the device with the feasible swap is moved to the new coalition, while the other stays in its current coalition and the new mutation is approved. Similarly, for scenario three, the swap of both devices is made and the new mutation is approved. For scenarios two and three, $\widetilde{\mathcal{P}}_{h-1}$ is generated and the corresponding association metrics are passed to Algorithm 1 to check for feasibility. If not feasible, then the mutation is rejected and $\mathcal{P}_h = \mathcal{P}_{h-1}$. If the partition, $\widetilde{\mathcal{P}}_{h-1}$, is feasible (i.e., problems **P7-1** and **P7-2** have solutions), the value functions of $\widetilde{\mathcal{P}}_{h-1}$ and of $\mathcal{P}_{h-1}$ are computed based on the definition in (5.4) and compared. If $\mathcal{V}(\widetilde{\mathcal{P}}_{h-1}) < \mathcal{V}(\mathcal{P}_{h-1})$, then $\mathcal{P}_h = \widetilde{\mathcal{P}}_{h-1}$ with probability 1. If $\mathcal{V}(\mathcal{P}_{h-1}) < \mathcal{V}(\widetilde{\mathcal{P}}_{h-1})$, then $\mathcal{P}_h = \mathcal{P}_{h-1}$ with probability $\sigma_h = \exp\{\delta/H\}$, and $\mathcal{P}_h = \widetilde{\mathcal{P}}_{h-1}$ with probability $1 - \sigma_h$, where $\delta = |\mathcal{V}(\mathcal{P}_{h-1}) - \mathcal{V}(\widetilde{\mathcal{P}}_{h-1})|$ and $H = \frac{H^0}{\log(h)}$ is a decaying design parameter corresponding to the number of times the partition update stage has been executed, where $H^0$ is the rate of decay. Further, let $\eta_0$ be the tolerance such that the *PAUSE* algorithm terminates if $H < \eta_0$. Note that, as we assume that *PAUSE* is ran by a centralized controller, we assume that the controller saves the latest partition and the corresponding optimal power and resource allocation matrices. Having introduced the basic operations of *PAUSE*, we can now formally present the algorithm as Algorithm 3.

### 5.4.2 Convergence and complexity analysis

As was shown, the *PAUSE* algorithm consists of two main stages: the initialization stage and the partition update stage. In the initialization stage, devices are sequentially moved from the extra coalition $C_{\mathcal{J}+1}$ to one of their feasible APs. The association matrices corresponding to each device move are generated and passed to Algorithm 1 to check feasibility. Thus, the initialization stage is composed of a series of Algorithm 1. To be able to analyze the complexity of the *PAUSE* algorithm, we need to analyze the complexity of Algorithm 1. In the feasibility check procedure, Algorithm 1 is used to solve the optimization problems **P7-1** and **P7-2** based on D.C. programming, hence, the complexity of Algorithm 1 can be evaluated using methods such as those presented in [43,94]. It should be noted that the complexity of Algorithm 1 depends on the underlying utilized optimization algorithm; thus, one algorithm can perform better than the other in terms of optimality and complexity [100,115]. For the following complexity analysis, we use the interior point method to solve the optimization problems. Other optimization methods can be used and the performances can be compared, however, this comparison is out of the scope of this work.

**Algorithm 2** Create the initial feasible partition $\mathcal{P}_0$

1: Initialize $\varrho$, $\vartheta = 1$, the partition $\mathcal{P}_0 = \{C_{1,0}, ..., C_{\mathcal{J}+1,0}\}$ such that $C_{j,0} = \emptyset$, for $j \leq \mathcal{J}$, and $C_{\mathcal{J}+1,0} = \mathcal{D}$. Initialize the association matrices $\mathrm{X}_0$ = [0], $\mathrm{Y}_0$ = [0], $\mathrm{S}_0$ = [0], $\mathrm{V}_0$ = [0], $\mathrm{Q}_{b,0}$ = [0], $\mathrm{Q}_{a,0}$ = [0]
2: **Repeat**
3: Let $i = 1$
4: Temp = $||C_{\mathcal{J}+1,0}||$
5: **While** $i \leq$ Temp:
6: Move the $i^{th}$ device from $C_{\mathcal{J}+1,0}$ to a randomly chosen coalition, $C_{j,0}$, in which the device has coverage, for $j \neq \mathcal{J}+1$.
7: Generate a corresponding temporary partition $\mathcal{P}^i_{0,\vartheta}$ and the corresponding association matrices $\mathrm{X}^i_{0,\vartheta}$ and $\mathrm{Y}^i_{0,\vartheta}$
8: Pass the temporary association matrices $\mathrm{X}^i_{0,\vartheta}$ and $\mathrm{Y}^i_{0,\vartheta}$ to Algorithm 1 to check for feasibility by finding solution points for problems **P7-1** and **P7-2**
**If** $\mathcal{P}^i_{0,\vartheta}$ = feasible
Move device $i$ from $C_{\mathcal{J}+1,0}$ to the coalition $C_{j,0}$
$\mathcal{P}_0 = \mathcal{P}^i_{0,\vartheta}$, $\mathrm{X}_0 = \mathbf{X}^i_{0,\vartheta}$, $\mathrm{Y}_0 = \mathbf{Y}^i_{0,\vartheta}$
$i = i + 1$
**Else**
Device $i$ stay in $C_{\mathcal{J}+1,0}$
$\mathcal{P}^i_{0,\vartheta} = \mathcal{P}_0$
$i = i + 1$
**End**
**EndWhile**
9: $\vartheta = \vartheta + 1$
10: Compute: $\rho = 1 - \exp\{-\kappa\vartheta\}$
**If** $C_{\mathcal{J}+1,0} = \emptyset$ or $\rho < \varrho$
Exit
**Else**
Temp = 2 and $i = 1$
Go to Step 5
**End**
11: **return** $\mathcal{P}_0$, $\mathrm{X}_0$, and $\mathrm{Y}_0$

**Algorithm 3** The *PAUSE* algorithm for solving problem **P2**

1: Initialization: Create the initial feasible partition $\mathcal{P}_0 = \{C_{1,0}, ..., C_{\mathcal{J}+1,0}\}$ using Algorithm 2. Initialize the matrices $\mathrm{X}_h = [0]$, $\mathrm{Y}_h = [0]$, $\mathrm{S}_h = [0]$, $\mathrm{V}_h = [0]$, $\mathrm{Q}_{b,h} = [0]$, $\mathrm{Q}_{a,h} = [0]$. Initialize $H^0$ and $h = 1$

2: **Repeat**

3: Randomly choose two coalitions from partition $\mathcal{P}_{h-1}$, say $C_{j,h-1}$ and $C_{j',h-1}$, for $j \neq j'$ and $0 < j \ \& \ j' \leq \mathcal{J} + 1$

4: Randomly choose a device from each of the chosen coalitions in Step 3, say $d_i$ from $C_{j,h-1}$ and $d_{i'}$ from $C_{j',h-1}$

5: Swap the chosen devices and obtain a temporary partition $\widetilde{\mathcal{P}}_{h-1}$. Swap is done if and only if it is feasible (i.e., the device is moved to a coalition where it has coverage)

6: Compute the AP association matrices $\widetilde{\mathbf{X}}_{h-1}$ and $\widetilde{\mathbf{Y}}_{h-1}$ as well as $\mathrm{X}_{h-1}$ and $\mathrm{Y}_{h-1}$ corresponding to partition $\mathcal{P}_{h-1}$

7: Pass the matrices $\widetilde{\mathbf{X}}_{h-1}$, $\widetilde{\mathbf{Y}}_{h-1}$, $\mathrm{X}_{h-1}$ and $\mathrm{Y}_{h-1}$ to Algorithm 1 to find the corresponding optimal resource and power allocation matrices

**If** $\widetilde{\mathcal{P}}_{h-1}$ is feasible

Compute corresponding value function for both $\widetilde{\mathcal{P}}_{h-1}$ and $\mathcal{P}_{h-1}$ denoted by $\mathcal{V}(\widetilde{\mathcal{P}}_{h-1})$ and $\mathcal{V}(\mathcal{P}_{h-1})$ respectively

**If** $|\mathcal{V}(\widetilde{\mathcal{P}}_{h-1}) - \mathcal{V}(\mathcal{P}_{h-1}))| > 0$

Compare $\mathcal{V}(\widetilde{\mathcal{P}}_{h-1})$ and $\mathcal{V}(\mathcal{P}_{h-1})$ **If** $\mathcal{V}(\widetilde{\mathcal{P}}_{h-1}) < \mathcal{V}(\mathcal{P}_{h-1})$

$\mathcal{P}_h = \widetilde{\mathcal{P}}_{h-1}$ with probability 1

**Else**

$\mathcal{P}_h = \mathcal{P}_{h-1}$ with probability $\sigma_h = \exp\{\delta/H\}$, and $\mathcal{P}_h = \widetilde{\mathcal{P}}_{h-1}$ with probability $1 - \sigma_h$

**End**

**Else**

$\mathcal{P}_h = \widetilde{\mathcal{P}}_{h-1}$ with probability 0.5, and $\mathcal{P}_h = \mathcal{P}_{h-1}$ with probability 0.5

**Exit**

**Else**

$\mathcal{P}_h = \mathcal{P}_{h-1}$ with probability 1

**End**

8: Save the value function, user association, resource allocation, and power assignment matrices of the updated partition

**If** $H < \eta_0$

**Exit**

**End**

9: $h = h + 1$

10: The optimal solution to problem **P2** and hence problem **P1** is equal to $\mathrm{X}_{h-1}$, $\mathrm{Y}_{h-1}$, $\mathrm{S}_{h-1}$, $\mathrm{V}_{h-1}$, $\mathrm{Q}_{b,h-1}$, $\mathrm{Q}_{a,_{h-1}}$

Let $\mathcal{D}_b$ and $\mathcal{D}_a$ denote the maximum number of devices associated with a BS and a DA respectively, i.e., $\mathcal{D}_b \triangleq \underset{j\in\Phi_b}{max}|C_j|$ and $\mathcal{D}_a \triangleq \underset{h\in\Phi_a}{max}|C_h|$. Thus, for problems **P7-1** and **P7-2**, we have a total of $2\mathcal{D}_bK\mathcal{B}+\mathcal{D}_b$ and $2\mathcal{D}_aN\mathcal{A}+\mathcal{D}_a$ decision variables respectively, where $\mathcal{D}_b+\mathcal{D}_a \leq \mathcal{D}$. Furthermore, there are $\mathcal{G}_b = K\mathcal{B}+2\mathcal{D}_bK\mathcal{B}+2\mathcal{D}_b\mathcal{B}$ and $\mathcal{G}_a = N\mathcal{A}+2\mathcal{D}_aN\mathcal{A}+2\mathcal{D}\mathcal{A}$ convex and linear constraints for problems **P7-1** and **P7-2** respectively. Accordingly, every iteration of Algorithm 1 to check the feasibility of a pair of generated assignment matrices $X$ and $Y$ has a computational complexity of the order $O((2\mathcal{D}_bK\mathcal{B}+2\mathcal{D}_aN\mathcal{A}+\mathcal{D}_a+\mathcal{D}_b)(\mathcal{G}_b+\mathcal{G}_a))$, where $O(\cdot)$ is the big-O notation.

Note that Algorithm 1 is used in both the initialization stage and in the partition update stage. In the initialization stage, Algorithm 1 is called at least $\mathcal{D}$ times. Algorithm 1 may be called more times in the the initialization stage depending on the control parameters $\kappa$ and $\varrho$. As discussed in the feasibility check stage of the PAUSE algorithm, the number of times Algorithm 1 is called decreases as a function of the number of re-executions $\vartheta$, at the rate $\exp\{-\kappa\vartheta\}$. Further, the initialization stage converges when $1-\exp\{-\kappa\vartheta\} < \varrho$. Therefore, the number of times the feasibility check is re-executed after the $\mathcal{D}^{th}$ time is given as $\vartheta = \frac{\log\left(\frac{1}{1-\varrho}\right)}{\kappa}$. Accordingly, the computational complexity of the initialization stage is of $(\vartheta+\mathcal{D})(O((2\mathcal{D}_bK\mathcal{B}+2\mathcal{D}_aN\mathcal{A}+\mathcal{D}_a+\mathcal{D}_b)(\mathcal{G}_b+\mathcal{G}_a)))$ iterations.

After the initialization stage is completed, the partition update stage commences. In the partition update stage, mutations of the initial partition are generated sequentially and passed to Algorithm 1 to check their feasibility. The number of times Algorithm 1 is called in the partition update stage is a function of the decay parameter $H = \frac{H^0}{\log(h)}$. Thus, as the partition update stage converges as $H$ approaches zero (i.e., $H < \eta_0$), the computational complexity of the partition update stage is of $(\exp\{\frac{H^0}{\eta_0}\})(O((2\mathcal{D}_bK\mathcal{B}+2\mathcal{D}_aN\mathcal{A}+\mathcal{D}_a+\mathcal{D}_b)(\mathcal{G}_b+\mathcal{G}_a)))$ iterations. By multiplying out the arguments within the parenthesis, we can see that the dominant terms are $4(\mathcal{D}_bK\mathcal{B})^2+4(\mathcal{D}_aN\mathcal{A})^2$. Consequently, the total computational complexity of the *PAUSE* algorithm in number of iterations is approximately given as

$$\mathcal{O} \approx (\vartheta+\mathcal{D}+\exp\{\frac{H^0}{\eta_0}\})(O(4(\mathcal{D}_bK\mathcal{B})^2+4(\mathcal{D}_aN\mathcal{A})^2)). \tag{5.15}$$

## 5.5 Simulation results and discussions

### 5.5.1 Algorithm performance examination

We evaluate the effectiveness of *PAUSE* via computer simulations with parameters listed in Table 5.1. Three different study cases are used to examine the impact of $\epsilon$, DA density $\lambda_a$, and the number of RCs $N$ on the performance of *PAUSE*. Studies are conducted on a square area

of 800 $m$ $x$ 800 $m$. It should be noted that, different from the simulations done in the results sub-sections of Chapters 3 and 4, in this part it is not possible to simulate a 200 $km$ $x$ 200 $km$ area due to hardware limitations. Due to the large number of variables that we solve for, the local PC used to run the algorithm takes a long time to compile, thus, in this section we provide a representative scaled down results for the purpose of proof of concept. Two BSs, each of circular coverage area with radius $\Lambda_b = 200m$, are located at $(200m, 200m)$ and $(600m, 600m)$ from the origin. For the initialization phase of the *PAUSE* algorithm, the devices are associated based on proximity and coverage. That is, a device can be associated with an AP if both of the following conditions are met: i) the physical distance between the device and the AP is less than the radius of coverage of that AP, and 2) the maximum transmission range of the device on any of the RBs/sub-channels of that AP is greater than or equal to the AP's coverage radius. As we consider a Rayleigh fading channel and path-loss propagation attenuation, the maximum transmission ranges of the $i^{th}$ device located at $d_i$ towards the $j^{th}$ BS, located at $b_j$, or towards the $h^{th}$ DA, located at $a_h$, are respectively given by

$$D_{i,j} = Q_{max}||d_i - b_j||^{-\alpha} g_{i,j}$$
$$D_{i,h} = Q_{max}||d_i - a_h||^{-\alpha} g_{i,h}$$

where $g_{i,j}$ and $g_{i,h}$ are the channel gain between the device and the AP.

It should be noted that, the objective from using this method of association in the initialization stage is to mimic the traditional max-RSS association mechanism. However, since in the initialization stage we use Algorithm 1 to optimize power consumption as well as resource utilization, the outcome of the initialization phase can be thought of as an optimized version of the max-RSS based user association. This serves as a baseline to compare the performance of the proposed *PAUSE* algorithm to as will be shown.

Table 5.1: Simulation parameters

| **Name** | **Value** | **Name** | **Value** |
|---|---|---|---|
| $w_b$ | 50 kHz | $\mu_1 = \mu_2$ | 5000 |
| $w_c$ | 100 kHz | $\mathcal{R}_{min}$ | 1 kbps |
| $w_u$ | 150 kHz | $e$ | 1e-3 |
| $\alpha$ | 4 | $H^0$ | 2 |
| $\Lambda_a$ | 40 m | $\eta_0$ | 0.5 |
| $\Lambda_b$ | 200 m | $\kappa$ | 1 |
| $Q_{max}$ | 500 mW | $\varrho$ | 0.5 |

#### 5.5.1.1 Cast Study I

The main objective is to study the impact of $\epsilon$ on the behavior the proposed *PAUSE* algorithm when solving problem **P2**. Recall that, as $\epsilon$ decreases, more weight emphasis is placed on maximizing the number of supported devices, while the minimization of the transmission power becomes of less importance. To study the impact of $\epsilon$ on the behavior of *PAUSE* when solving **P2**, $\epsilon \in \{0.0008, 0.001, 0.0014, 0.002, 0.0041\}$, are chosen such that the condition set in (5.2) is met. The network is made up of ten DAs, each of radius $\Lambda_a = 40\ m$, randomly deployed across the 2D plane within the simulate square area. Each DA has $N = 10$ RCs with $\omega_c = 100\ kHz$. There are $\mathcal{D} = 240$ devices randomly distributed over the region. The initial partition successfully serves 84 devices with an average per device transmission power of 94.7626 $mW$ for all $\epsilon$ values. Starting from this point, the partition update stage of the *PAUSE* algorithm is used to optimize the performance for different $\epsilon$ values.

The results after convergence for different $\epsilon$ values are given in Table 5.2. It is observed that increasing $\epsilon$ prioritizes reducing power consumption over increasing the number of supported devices. For instance, at $\epsilon = 0.0008$, the algorithm successfully serves 9 more devices compared to the initial point, while the per device average transmission power increases from 94.7626 $mW$ to 113.9413 $mW$. On the other hand, at $\epsilon = 0.0041$, only 2 more devices are served compared to the initial point, however, the transmission power decreases to 77.9303 $mW$. From problem **P2**, it can be seen that, as $\epsilon$ approaches zero, the problem simplifies to a maximization problem of the number of supported devices and ignores the minimization of transmission power. This is supported by the results presented in Figure 5.2, where we set $\epsilon = 0$ and the algorithm converges to the absolute maximum possible number of 95 devices. For the results shown in Figure 5.2, we harden the stopping criteria as the objective is to push the algorithm to the limit and see the long term behaviour. From the shown results, we can conclude that there is a maximum number of devices that can be successfully supported. The results suggest that, for all $\epsilon$ values, not all the devices can be accommodated due to the interference which limits the possibility of supporting more devices at the required minimum data rate.

Table 5.2: Impact of $\epsilon$ on the behaviour of the *PAUSE* algorithm

| **Epsilon ($\epsilon$)** | **Value of partition** | **No. of assoc. devices** | **Avg. trans. power (mW)** | **Value of opt. partition** | **Value diff.** |
|---|---|---|---|---|---|
| **0.0041** | 44.8029 | 86 | 77.9303 | 58.1692 | 13.3663 |
| **0.0020** | 64.8795 | 88 | 93.7671 | 71.3210 | 6.4415 |
| **0.0014** | 70.6156 | 89 | 108.1003 | 74.9077 | 4.2921 |
| **0.0010** | 74.4397 | 90 | 112.1331 | 79.8180 | 5.3783 |
| **0.0008** | 76.3518 | 93 | 113.9413 | 84.4484 | 8.0966 |

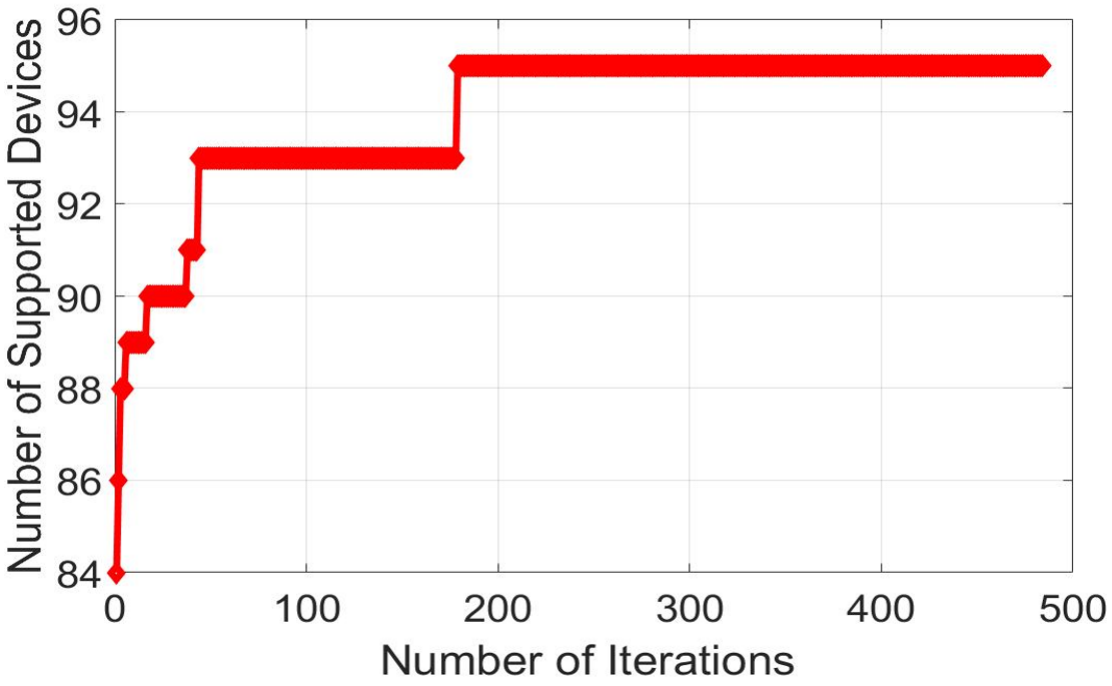


Figure 5.2: The maximum possible number of devices that can be supported at $\epsilon = 0$.

#### 5.5.1.2 Case study II

To study the impact of DA density as a design parameter, we test the network when it has 4, 8, 16, and 32 DAs that are randomly deployed each with coverage area of radius $\Lambda_a = 40\ m$. Each DA has 10 RCs each of BW 100 $kHz$. A total of 280 devices are randomly deployed. For all DA densities, $\epsilon = 0.0014$. Table 5.3 shows the total number of supported devices and their distribution between BSs and DAs (the $3^{rd}$ column), the average transmission power per device (the $4^{th}$ column) after the initialization phase, and the resultant values (in the $5^{th}$ and $6^{th}$ columns) after the *PAUSE* algorithm converges. The following observations can be made:

Table 5.3: Impact of DA density

| **DAs used** | **Available RB + RC** | **Supported (BS, DA, total)** | **Ave. trans. power (mW)** | **Opt. No. of (BS, DA, total)** | **Opt. Device ave. trans. power (mW)** |
|---|---|---|---|---|---|
| **4** | 40+40 | (31 , 35, 66) | 113.5642 | (30, 36, 66) | 113.2334 |
| **8** | 40+80 | (28, 55, 83) | 109.3373 | (29, 64, 93) | 131.7634 |
| **16** | 40+160 | (25, 98, 123) | 87.1789 | (25, 105, 130) | 97.9106 |
| **32** | 40+320 | (15, 160, 175) | 70.1782 | (16, 174, 190) | 67.2939 |

i) Increasing the number of DAs increases the amount of available resources and hence the network can support more devices, however, the increase is non-linear due to increased interference; ii) as the DA number increases, more devices are associated with them, which reduces the load on the BSs; iii) there is an optimal number of DAs that results in the best

overall network coverage. This is evident when the number of DAs increases from 4 to 8 to 16 which results in an increase in the per device average transmission power compared to the initial values. This is because *PAUSE* is associating devices that may be far away leading to an increase in the average transmission power. However, when 32 DAs are used, network coverage is maximized and *PAUSE* now optimizes power by including the devices that best maximize the number of satisfied users at the minimum transmission power possible; and iv) more DAs leads to an increase in the number of supported devices, yet the network is unable to support all the 280 devices. As the number of DAs increases, the amount of resources allocated for every DA decreases, leading to a lower average achievable data rate per DA. As the QoS requirements of the devices using two-hop routes depends on the link rates of the two-hop path, lower DA uplink data rate limits the number of devices a DA can support. This result highlights the dependency between the two-hop route and the available resources; there is an optimal number of DAs for maximal cost effectiveness.

#### 5.5.1.3 Case study III

The network has 20 randomly deployed DAs. To study the impact of varying the number of available RCs at each DA, we set the number of RCs, $N$, to 10, 15, 20, and 30 each of BW 100 $kHz$. Accordingly, the total number of available uplink resources (RBs+RCs) is in 240, 340, 440, and 640 respectively. A total of 350 devices are randomly deployed. Results of this case study are shown in Table 5.4, with $\epsilon$ = 0.0014. Based on the results in Table 5.4, few observations can be made: i) As $N$ increases, the number of devices that can be successfully supported increases, and the average per device transmission power decreases. It is expected, as more RUs are available for the devices and the devices are more dispersed across the available RUs, leading to a decrease in interference and in the transmission power needed to achieve the required QoS; ii) a DA can only serve the devices located within its vicinity; hence, increasing the number of RCs does not always lead to an increase in the number of supported devices. For example, increasing the RCs from 20 to 30 does not lead to an increase in the maximum number of supported devices after convergence; iii) increasing the number of RCs might lead to an improvement in the average transmission power of the devices. This is because increasing the number of RCs provides more options for the disperse of the devices across the RCs. Using *PAUSE* algorithm, devices are assigned the RCs that best improve their transmission power; iv) the increase of the number of supported devices is non-linear with respect to the increase in the number of RCs due to interference and competition; and v) increasing the number of RCs may lead to a slight increase in the per average transmission power of the devices which can be

avoided by choosing different $\epsilon$ values to put emphasis on the desired objective. In conclusion, the number of DAs and the number of RCs are important network design factors that should be jointly optimized to support the largest number of devices at the lowest transmission power.

Table 5.4: Impact of varying $N$ on the behaviour of the *PAUSE* algorithm

| **No. of RCs** | **No. of Served Devices** | **Per device ave. trans. power (mW)** | **Opt. No. of Served Devices** | **Per device ave. trans. power (mW)** |
|---|---|---|---|---|
| **10** | 96 | 134.4565 | 106 | 136.1226 |
| **15** | 142 | 105.7782 | 164 | 122.7326 |
| **20** | 255 | 92.498 | 308 | 82.2356 |
| **30** | 275 | 76.9939 | 308 | 72.449 |

### 5.5.2 Performance comparison

For performance comparison, we consider three main baseline schemes, namely: random AP association scheme, max-RSS AP association scheme [29], and optimized max-RSS AP association scheme. In the random AP association scheme, each device is first randomly associated with one of its reachable APs. Algorithm 1 is then used to find the optimal resource assignments and power allocations for the current associations. In the max-RSS scheme, each device is associated with the AP to which it has the best channel conditions, however, resource assignment and power allocation are done randomly. On the other hand, the optimized max-RSS based scheme works in a way similar to the max-RSS scheme, except once the devices are associated, Algorithm 1 is then used to find the optimal resource assignments and power allocations for the current partition.

#### 5.5.2.1 Performance comparison at different DA densities

For the same 800 $m$ $x$ 800 $m$ square area, we simulate a network that contains the two BSs, each with $K$ = 20 RBs, and a total of 280 devices to be served. We vary the number of DAs in the network from 4 to 32 DAs that are randomly deployed. Each DA has 15 sub-channels and a coverage area of $\Lambda_a = 40m$. In Figure 5.3, we show the impact of varying the DA density on the average transmission power per device as well as the number of successfully served devices. Each result is an average of 100 Monte Carlo runs.

As can be seen from Figure 5.3, compared to all the baseline algorithms, the proposed algorithm significantly improves both the number of successfully served devices and the transmission energy per device. For instance, at 32 deployed DAs, the proposed algorithm serves 108 more devices than the random approach, and 53 more compared to optimized max-RSS. At the same time, it achieves 40% reduction in energy consumption compared to the random approach and

performs similarly to optimized max-RSS. These results demonstrate that the proposed algorithm achieves the maximum number of successfully served devices at the minimum energy consumption.

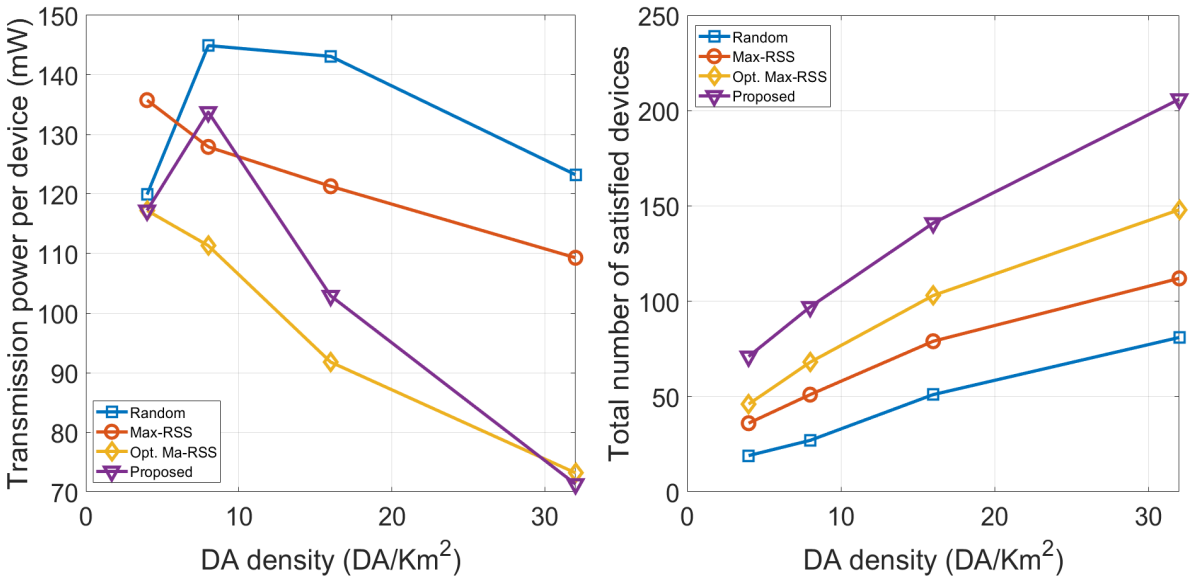


Figure 5.3: Impact of changing DA density on the average energy consumption per device (left) and on the number of satisfied served-devices (right)

#### 5.5.2.2 Performance comparison at different numbers of sub-channels per DA

We simulate the same network with 280 devices, the two BSs, and 30 randomly deployed DAs. The number of sub-channels per DA is varied from 10 to 25 in a step size of 5. Figure 5.4 shows the impact of varying the number of sub-channels per DA on the average transmission power per device as well as the number of successfully served devices. Each result is an average of 100 Monte Carlo runs. We choose to use 30 DAs for this study as to eliminate the impact of DA random positioning by forcing the system into the plateau region indicated in the results shown in Figure 5.3. From Figure 5.4, increasing the number of sub-channels, increases the number of supported devices. However, as we are operating in the plateau region, due to the high number of DAs deployed, the increase in the number of supported devices is not as significant as shown in Figure 5.3. On the other hand, increasing the number of sub-channels significantly improves the energy consumption per device. This is expected, as devices have more sub-channels to choose from, resulting in a lower number of interfering devices per sub-channel. Figure 5.4 also shows that the proposed algorithm outperforms all other examined baseline schemes in both power consumption and number of successfully served devices. At 30 DAs and 25 sub-channel per DA, the proposed algorithm can successfully serve all the 280 devices in the network.

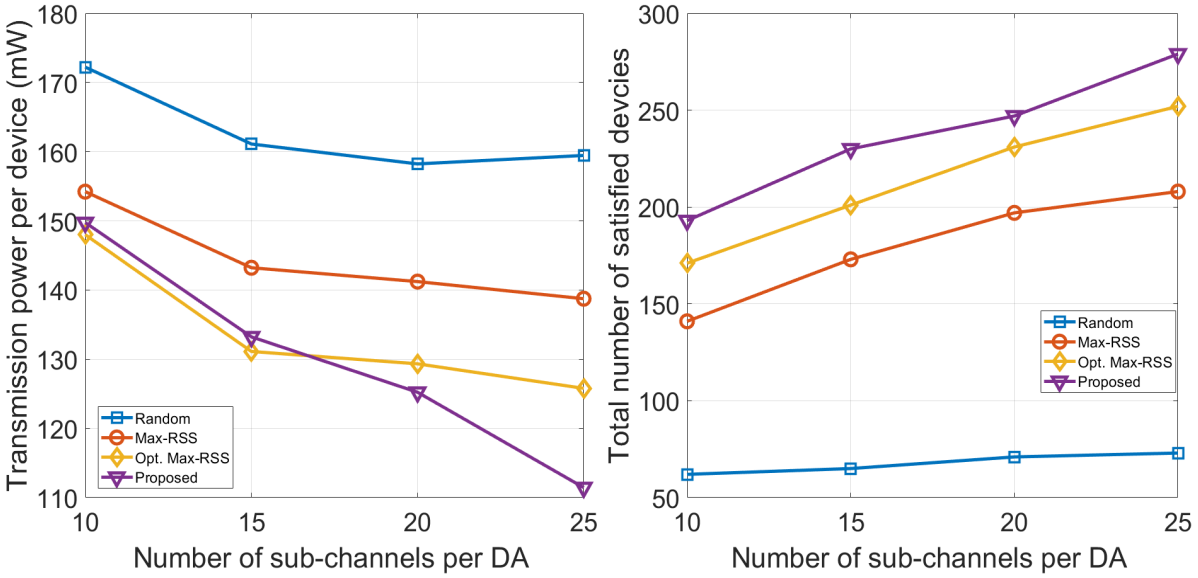


Figure 5.4: Impact of changing the number of sub-channels per DA on the average energy consumption per device (left) and on the number of satisfied served-devices (right)

## 5.6 Summary

In this Chapter, we study a two-hop DA infused cellular network to support a large number of battery power IoT devices. The objective is to maximize the number of satisfied devices while minimizing their energy consumption and achieving their desired minimum data rate. Complementing the cellular network with a layer of low-cost yet powerful aggregator nodes to relay data from the devices in a two-hop manner is an effective approach. However, the aggregator nodes are cellular devices themselves and thus use radio resources at the BSs, leading to a dependency that requires the optimization of the number of aggregators to be deployed given the available resources and the number of devices to be supported. We formulate the joint user association, power consumption minimization, and resource utility maximization problem as a MINLP multi-objective optimization problem from the view point of both devices and the network. We propose a novel algorithm, *PAUSE*, based on coalition formation games and D.C programming to simplify the problem into continuous convex non-linear sub-problems that can be solved using traditional optimization methods in polynomial time. Numerical results are presented to study the impact of summation weighted factor $\epsilon$, DA density $\lambda_a$, and the number of DA sub-channels $N$ on the algorithm performance and demonstrate the benefits of the proposed two-hop DA infused cellular network in supporting future massive cellular IoT applications. The performance of the proposed algorithm is compared to three baseline AP association schemes and is shown to outperform all of them in terms of the number of successfully served devices and energy consumption.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

The objective of this research is to develop a channel access mechanism to enable future anticipated massive cellular IoT applications that will have stringent QoS requirements. To achieve the objective, we divide the work into the stages described below:

- Analyze the performance of the current channel access mechanism and identify its limitations;
- Based on the identified limitations, develop a cost-effective, QoS aware, energy and radio resource efficient approach to improve the performance of the cellular network to support massive number of IoT devices;
- Analyze the performance of the proposed solution and identify its advantages and disadvantages;
- Consider a realistic network scenario and study the applicability of the proposed solution.

At the first stage, we study the RACH procedure, the channel access mechanism used in traditional cellular networks, and model its performance. In RACH, a device randomly chooses a preamble from a fixed set of preambles and transmits it to its desired BS. As the set of preambles are shared by all the devices in the network, the same preamble can be chosen by more than one device and transmitted at the same time, leading to interference and possible denial of access. As the preambles are used by all BSs, interference consists of two components: intra-cell interference and inter-cell interference. Characterizing and modeling the interference components is crucial to analyzing the performance of the RACH procedure. Yet, due to the complexity added by the random distances and locations of interfering devices on a preamble, traditional modeling techniques are limited. Accordingly, in our work, we proposed to use stochastic geometry to derive statistical models for the association success probability, taking

into account the spatial characteristics of the interfering devices. The derived models are shown to be accurate by means of computer simulations and numerical results. The models are used to study the impact of different variables such as the density of the devices, density of BSs, number of preambles, and FPC factor, on the association success probability. We draw the following conclusions:

- The RACH procedure acts as a bottleneck when the number of contending devices increases;
- As the number of devices increases, the access delay increases and can reach an unacceptable level for some IoT applications that are delay intolerant;
- Power control is a plausible way of improving the performance of the RACH procedure under massive number of device; however, the improvement comes at the cost of elevated levels of transmission power and energy consumption;
- Our work results in novel mathematical approaches to model complex uplink transmission paradigms that are anticipated to constitute a major portion of future wireless communication traffic.

These observations suggest the inadequacy of the current association mechanism of the cellular network when supporting large-scale IoT applications and necessitates finding novel channel access mechanisms to be part of future 5G cellular networks. The main limitations are due to the high contention at the RACH stage that leads to excessive delays, and the long transmission distances between devices and BSs that result in high energy consumption. Hence, in the second stage of the work, we propose to use data aggregation as a means to enable massive access on cellular networks. Data aggregation has the advantage of providing energy efficient communication as the DAs are often at a closer proximity to the devices than the cellular BSs. It also improves resource utilization with increased payload in each larger packet. Furthermore, by grouping devices into clusters and allowing cluster heads to be the only devices contending over the RACH preambles, access delay is greatly reduced.

Given these advantages, in the second part of the Ph.D research, we propose a two-hop network architecture that consists of a layer of cellular BSs overlaid with a layer of DAs. Devices transmit their data by first associating with their closest DA which then relays the data to the closest BS. To test the performance of the proposed architecture, we borrow tools from stochastic geometry and queuing theory, where we derive accurate models for the achievable data rate, association success probability, and end-to-end system delay. We also explore methods by which the two-hop architecture can be optimized to improve the overall network performance.

Besides exploring the advantages of clustering and data aggregation, our approach is unique as it considers NOMA for in-cluster transmissions to improve resource utilization and increase the number of supportable devices per-cluster. The objective is to use the minimum number of DAs that provide sufficient coverage for the devices, while limiting the contention over the RACH preambles among the DAs. NOMA allows multiple devices to be multiplexed on the same channel at the cost of controlled interference and slight increase in transmission power. We explore the potentials brought by combining NOMA and data aggregation for the purpose of supporting massive cellular IoT applications. Based on our models and numerical results, the following conclusions are made:

- We showcase the advantages of two-hop NOMA-enabled network architecture as compared to single-hop networks;
- We develop novel mathematical models using stochastic geometry and queuing theory that can be used to analyze the uplink performance of two-hop networks with NOMA-enabled links;
- Based on the proposed two-hop network architecture, we conclude that DA density as well as network SINR thresholds are two critical parameters that highly impact network performance and thus should be carefully designed;
- Compared to a traditional two-hop OMA based architecture, the proposed NOMA architecture provides better scheduling and coverage probability and supports a larger number of devices per transmission frame;
- NOMA results in a lower average end-to-end system delay compared to OMA;
- the advantages of NOMA come at the cost of increased energy consumption on the device's side.

After studying and showing the advantages of two-hop network architecture in efficiently supporting a massive number of devices, in the third part of the research, we consider more realistic network scenario where devices have the choice between directly connecting to BSs or transmitting their data in a two-hop fashion via DAs. In this scenario, to which AP a device connects is crucial and impacts the overall performance of the network. Thus, in this part, we consider optimized user association and resource allocation to enable energy efficient communication while catering for strict QoS requirements. The objective is to support the largest number of devices at their required data rate, while efficiently utilizing network radio

resources and minimizing energy consumption. The formulated problem is non-linear mixed integer problem that cannot be solved using convex approaches. Consequently, we propose the *PAUSE* algorithm to accurately solve the problem in polynomial time. In *PAUSE*, with the help of coalition game theory and stochastic geometry, the problem is divided into simpler sub-problems that are then transformed into a difference between two concave functions programming that can be solved by means of iterative approaches. The efficiency of the proposed algorithm is presented.

## 6.2 Future research direction

In this research, we present rigorous analysis of the anticipated massive access problem on future cellular networks due to future massive cellular IoT applications. The results from the above mentioned three stages pave the way for further investigations in the field of data aggregation and clustering as means for enabling massive channel access on future cellular networks. We also present some novel modeling approaches that consider various networking factors making them more comprehensive and accurate. The models can be used to study other parameters and draw more conclusions on future network designs. We also explore the potentials of NOMA as a way of increasing the number of supported devices while being resource efficient. We finally touch on the user association problem and present a new approach to solve the joint optimization problem of user association and radio resource utilization while taking into account the QoS requirement of the devices.

The detailed results show the potentials of data aggregation and further investigation is needed under more general scenarios. For example, in this work, we only consider single and two-hop transmission; however, multi-hop transmission might be a better approach for some applications especially those involving nodes placed at inaccessible locations such as under water. Considering multi-hop scenarios might be challenging when modeling the performance, yet stochastic geometry can be used to develop accurate statistical models.

Furthermore, in solving the user association problem, we do not consider NOMA due to its complexity. Nonetheless, as the second stage of the work has shown the advantages brought on by NOMA, it might be of great interest to study the user association problem in NOMA-enabled scenarios.

Last, in both parts two and three of this work, we did not consider the impact of RACH on the overall performance. As RACH is expected to be the main channel access mechanism for future cellular networks, considering its impact when analyzing multi-hop network architecture is crucial and is yet to be studied.

# Extracted Publications

1. Hesham G. Moussa and Weihua Zhuang. Access point association in uplink two-hop cellular IoT networks with data aggregators. IEEE Internet Things J., pages 1–1, Mar. 2020.
2. Hesham G. Moussa and Weihua Zhuang. Energy- and delay-aware two-hop NOMA-enabled massive cellular IoT communications. IEEE Internet Things J., 7(1):558–569, Jan. 2020.
3. Hesham G. Moussa and Weihua Zhuang. RACH performance analysis for large-scale cellular IoT applications. IEEE Internet Things J., 6(2):3364–3372, Apr. 2019.

# Bibliography


[1] A. Ahmed Abbasi and M. Younis. A survey on clustering algorithms for wireless sensor networks. *Comput. Commun.*, 30(14-15):2826–2841, Oct. 2007.

[2] M Mehdi Afsar and Mohammad-H Tayarani-N. Clustering in sensor networks: A literature survey. *J. of Net. and Comput. Appl.*, 46:198–226, 2014.

[3] A. Ahmed, L. M. Boulahia, and D. Gaïti. Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification. *IEEE Commun. Surveys & Tuts.*, 16(2):776–811, Aug. 2014.

[4] Z. Q. Al-Abbasi and D. K. C. So. User-pairing based non-orthogonal multiple access (NOMA) system. In *Proc. IEEE VTC*, pages 1–5, May 2016.

[5] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli. Uplink non-orthogonal multiple access for 5G wireless networks. In *Proc. IEEE ISWCS*, pages 781–785, Aug. 2014.

[6] Tanweer Alam. A reliable communication framework and its use in internet of things (IoT). *Proc. Int. J. of Scientific. Research in Comp. Science, Eng. and Info. Tech. (IJSRCSEIT)*, 3(5):450–456, 2018.

[7] O. L. Alcaraz López, H. Alves, P. H. Juliano Nardelli, and M. Latva-aho. Aggregation and resource scheduling in machine-type communication networks: A stochastic geometry approach. *IEEE Trans. Wireless Commun.*, 17(7):4750–4765, Jul. 2018.

[8] M. S. Ali, E. Hossain, and D. I. Kim. LTE/LTE-A random access for massive machine-type communications in smart cities. *IEEE Commun. Mag.*, 55(1):76–83, Jan... 2017.

[9] J. G. Andrews, F. Baccelli, and R. K. Ganti. A tractable approach to coverage and rate in cellular networks. *IEEE Trans. on Commun.*, 59(11):3122–3134, Nov. 2011.

[10] J. G. Andrews, R. K. Ganti, M. Haenggi, N. Jindal, and S. Weber. A primer on spatial modeling and analysis in wireless networks. *IEEE Commun. Mag.*, 48(11):156–163, 2010.

[11] J. G. Andrews, A. K. Gupta, and H. S. Dhillon. A primer on cellular network analysis using stochastic geometry. *arXiv preprint arXiv:1604.03183*, 2016.

[12] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon. An overview of load balancing in hetnets: old myths and open problems. *IEEE Wireless Commun.*, 21(2):18–25, Apr. 2014.

[13] W. Bao and B. Liang. Structured spectrum allocation and user association in heterogeneous cellular networks. In *Proc. IEEE INFOCOM*, pages 1069–1077, Apr. 2014.

[14] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura. Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials. In *Proc. Int. Conf. on Wireless Net. and Mobile Commun. (WINCOM)*, pages 1–6, Oct. 2015.

[15] B. Błaszczyszyn and D. Yogeshwaran. Clustering comparison of point processes, with applications to random geometric models. In *Stochastic Geometry, Spatial Statistics and Random Fields*, volume 2120, pages 31–71. Lecture Notes in Mathematics. Cham, Switzerland: Springer-Verlag, 2015.

[16] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[17] B. Błaszczyszyn, M. K. Karray, and H. P. Keeler. Using poisson processes to model lattice cellular networks. In *Proc. IEEE INFOCOM*, pages 773–781, 2013.

[18] V. Chandrasekhar and J. G. Andrews. Uplink capacity and interference avoidance for two-tier femtocell networks. *IEEE Trans. Wireless Commun.*, 8(7):3498–3509, Jul. 2009.

[19] E. Che, H. D. Tuan, and H. H. Nguyen. Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks. *IEEE Trans. on Wireless Commun.*, 13(10):5481–5495, Oct. 2014.

[20] Hong Chen and David D Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, volume 46. Springer Science & Business Media, 2013.

[21] S. Chen, R. Ma, H. H. Chen, H. Zhang, W. Meng, and J. Liu. Machine-to-machine communications in ultra-dense networks–A survey. *IEEE Commun. Surveys & Tuts.*, 19(3):1478–1503, Mar. 2017.

[22] W. C. Cheung, T. Q. S. Quek, and M. Kountouris. Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks. *IEEE J.on Selected Areas in Commun.*, 30(3):561–574, Apr. 2012.

[23] W. H. Chin, Z. Fan, and R. Haines. Emerging technologies and research challenges for 5G wireless networks. *IEEE Wireless Commun.*, 21(2):106–112, Apr. 2014.

[24] J. Choi. NOMA-based random access with multichannel ALOHA. *IEEE J.on Selected Areas in Commun.*, 35(12):2736–2743, Dec. 2017.

[25] Q. H. Chu, J. M. Conrat, and J. C. Cousin. Propagation path loss models for LTE-advanced urban relaying systems. In *Proc. IEEE APSURSI*, pages 2797–2800, Jul. 2011.

[26] X. Chu, D. Lopez-Perez, Y. Yang, and F. Gunnarsson. *Heterogeneous cellular networks: Theory, simulation and deployment*. Cambridge University Press, New York, NY, USA, 2013.

[27] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.*, 53(9):74–81, Sep. 2015.

[28] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca. Allocation of control resources for machine-to-machine and human-to-human communications over LTE/LTE-A networks. *IEEE Internet Things J.*, 3(3):366–377, Jun. 2016.

[29] H. S. Dhillon, R. K. Ganti, and J. G. Andrews. Load-aware modeling and analysis of heterogeneous cellular networks. *IEEE Trans. Wireless Commun.*, 12(4):1666–1677, Apr. 2013.

[30] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews. Modeling and analysis of K-tier downlink heterogeneous cellular networks. *IEEE J.on Selected Areas in Commun.*, 30(3):550–560, 2012.

[31] H. S. Dhillon, H. Huang, and H. Viswanathan. Wide-area wireless communication challenges for the Internet of Things. *IEEE Commun. Mag.*, 55(2):168–174, Feb. 2017.

[32] M. Di Renzo and P. Guan. Stochastic geometry modeling and system-level analysis of uplink heterogeneous cellular networks with multi-antenna base stations. *IEEE Trans. on Commun.*, 64(6):2453–2476, Jun. 2016.

[33] Z. Ding, P. Fan, and H. V. Poor. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans. on Veh. Tech.*, 65(8):6010–6023, Aug. 2016.

[34] Z. Ding, Z. Yang, P. Fan, and H. V. Poor. On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Sig. Processing Letters*, 21(12):1501–1505, Dec. 2014.

[35] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong. D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks. *IEEE Trans. Veh. Technol.*, 65(12):9847–9861, Dec. 2016.

[36] M. K. Elhattab, M. M. Elmesalawy, and I. I. Ibrahim. Opportunistic device association for heterogeneous cellular networks with H2H/IoT co-existence under QoS guarantee. *IEEE Internet Things J.*, 4(5):1360–1369, Oct. 2017.

[37] H. ElSawy and E. Hossain. On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control. *IEEE Trans. Wireless Commun.*, 13(8):4454–4469, Aug. 2014.

[38] H. ElSawy, E. Hossain, and M. Haenggi. Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey. *IEEE Commun. Surveys & Tuts.*, 15(3):996–1019, Jun. 2013.

[39] H. ElSawy, A. Sultan-Salem, M. S. Alouini, and M. Z. Win. Modeling and Analysis of Cellular Networks Using Stochastic Geometry: A Tutorial. *IEEE Commun. Surveys & Tuts.*, 19(1):167–203, Nov. 2017.

[40] ETSI. LTE; evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (3GPP TS 36.211 version 14.2.0 Release 14). *TS 36.211 version 14.2.0*, 2017.

[41] M Ferraro and L Zaninetti. On the statistics of area size in two-dimensional thick Voronoi diagrams. *Physica A: Stat. Mech. and its Appl.*, 391(20):4575–4582, Jul. 2012.

[42] Mo-Han Fong, Puneet K Jain, and Hyung-Nam Choi. Extended access barring, Feb. 2016. US Patent 9,264,979.

[43] Anders Forsgren, Philip E Gill, and Margaret H Wright. Interior methods for nonlinear optimization. *SIAM review*, 44(4):525–597, 2002.

[44] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy. Energy and delay analysis of LTE-advanced RACH performance under MTC overload. In *Proc. IEEE Globecom Workshops*, pages 1632–1637, Dec. 2012.

[45] M. Gharbieh, H. ElSawy, A. Bader, and M. S. Alouini. Tractable stochastic geometry model for IoT access in LTE networks. In *Proc. IEEE Globecom*, pages 1–7, Dec. 2016.

[46] M. Gharbieh, H. ElSawy, A. Bader, and M. S. Alouini. Spatiotemporal stochastic modeling of IoT enabled cellular networks: Scalability and stability analysis. *IEEE Trans. Commun.*, 65(8):3585–3600, Aug. 2017.

[47] J. Ghimire and C. Rosenberg. Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach. *IEEE Trans. on Wireless Commun.*, 12(3):1340–1351, Mar. 2013.

[48] J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu. Massive machine type communication with data aggregation and resource scheduling. *IEEE Trans. Commun.*, 65(9):4012–4026, Sep. 2017.

[49] I. Guvenc. Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination. *IEEE Commun. Letters*, 15(10):1084–1087, Oct. 2011.

[50] V. N. Ha and L. B. Le. Distributed base station association and power control for heterogeneous cellular networks. *IEEE Trans. on Veh. Tech*, 63(1):282–296, Jan.. 2014.

[51] M. Haenggi. *Stochastic Geometry for Wireless Networks.* Cambridge University Press, New York, NY, USA, 1st edition, 2012.

[52] M. Hasan, E. Hossain, and D. Niyato. Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches. *IEEE Commun. Mag.*, 51(6):86–93, Jun. 2013.

[53] G. Hattab and D. Cabric. Performance analysis of uplink cellular IoT using different deployments of data aggregators. In *Proc. IEEE Globecom*, pages 1–6, Dec. 2018.

[54] Ghaith Hattab and Danijela Cabric. Energy-efficient massive IoT shared spectrum access over UAV-enabled cellular networks. *arXiv preprint arXiv:1808.08006*, 2018.

[55] M. Hong and Z. Luo. Distributed linear precoder optimization and base station selection for an uplink heterogeneous network. *IEEE Trans. on Sig. Processing*, 61(12):3214–3228, Jun. 2013.

[56] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser. Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective. *IEEE Wireless Commun.*, 21(3):118–127, Jun. 2014.

[57] R. Q. Hu and Y. Qian. *Heterogeneous cellular networks.* John Wiley & Sons, 2013.

[58] H. Ibrahim, W. Bao, and U. T. Nguyen. Data rate utility analysis for uplink two-hop internet of things networks. *IEEE Internet Things J.*, 6(2):3601–3619, Apr. 2019.

[59] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan. Random access analysis for massive IoT networks under a new spatio-temporal model: A stochastic geometry approach. *IEEE Trans. on Commun.*, pages 1–16, Jul. 2018.

[60] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek. Analyzing random access collisions in massive IoT networks. *IEEE Trans. on Wireless Commun.*, 17(10):6853–6870, Oct. 2018.

[61] H. Jo, Y. J. Sang, P. Xia, and J. G. Andrews. Outage probability for heterogeneous cellular networks with biased cell association. In *Proc. IEEE Globecom)*, pages 1–5, Dec. 2011.

[62] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews. Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis. *IEEE Trans. Wireless Commun.*, 11(10):3484–3495, Oct. 2012.

[63] B. Khamidehi, A. Rahmati, and M. Sabbaghian. Joint sub-channel assignment and power allocation in heterogeneous networks: An efficient optimization method. *IEEE Commun. Letters*, 20(12):2490–2493, Dec. 2016.

[64] D. M. Kim, R. B. Sorensen, K. Mahmood, O. N. Osterbo, A. Zanella, and P. Popovski. Data aggregation and packet bundling of uplink small packets for monitoring applications in LTE. *IEEE Network*, 31(6):32–38, 2017.

[65] J. Kim, H. Lee, D. M. Kim, and S. Kim. Delay performance of two-stage access in cellular Internet-of-Things networks. *IEEE Trans. Veh. Technol.*, 67(4):3521–3533, Apr. 2018.

[66] T. Kwon and J. M. Cioffi. Random deployment of data collectors for serving randomly-located sensors. *IEEE Trans. Wireless Commun.*, 12(6):2556–2565, Jun. 2013.

[67] A. Larmo and R. Susitaival. RAN overload control for machine type communications in LTE. In *Proc. IEEE Globecom Workshops*, pages 1626–1631, Dec. 2012.

[68] S. Lasaulce and H. Tembine. *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. Academic Press, 1st edition, 2011.

[69] 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 Commun. Surveys & Tuts.*, 16(1):4–16, Dec. 2013.

[70] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner. Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks. In *Proc. IEEE ICC*, pages 1–6, May 2016.

[71] A. Li, Y. Lan, X. Chen, and H. Jiang. Non-orthogonal multiple access (NOMA) for future downlink radio access of 5G. *China Commun.*, 12(Supplement):28–37, Dec. 2015.

[72] Duan Li and Xiaoling Sun. *Nonlinear integer programming*, volume 84. Springer, Berlin, Heidelberg, 2006.

[73] W. Li, Q. Du, L. Liu, P. Ren, Y. Wang, and L. Sun. Dynamic allocation of RACH resource for clustered M2M communications in LTE networks. In *Proc. Int. Conf. on Ident., Info., and Knowledge in the Internet of Things (IIKI)*, pages 140–145, 2015.

[74] L. Liang, L. Xu, B. Cao, and Y. Jia. A cluster-based congestion-mitigating access scheme for massive M2M communications in internet of things. *IEEE Internet Things J.*, 5(3):2200–2211, Jun. 2018.

[75] G. Y. Lin, S. R. Chang, and H. Y. Wei. Estimation and adaptation for bursty LTE random access. *IEEE Trans. Veh. Technol.*, 65(4):2560–2577, Apr. 2016.

[76] T. Lin, C. Lee, J. Cheng, and W. Chen. PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks. *IEEE Trans. Veh. Technol.*, 63(5):2467–2472, Jun. 2014.

[77] Y. Lin, W. Bao, W. Yu, and B. Liang. Optimizing user association and spectrum allocation in HetNets: A utility perspective. *IEEE J.on Selected Areas in Commun.*, 33(6):1025–1039, 2015.

[78] D. Liu, Y. Chen, K. K. Chai, and T. Zhang. Joint uplink and downlink user association for energy-efficient HetNets using nash bargaining solution. In *Proc. IEEE VTC*, pages 1–5, May 2014.

[79] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan. Opportunistic user association for multi-service HetNets using nash bargaining solution. *IEEE Commun. Letters*, 18(3):463–466, Mar. 2014.

[80] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. Wong, R. Schober, and L. Hanzo. User association in 5G networks: A survey and an outlook. *IEEE Commun. surveys & Tuts.*, 18(2):1018–1044, Secondquarter 2016.

[81] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan. Distributed energy efficient fair user association in massive MIMO Enabled HetNets. *IEEE Commun. Letters*, 19(10):1770–1773, Oct. 2015.

[82] D. Lopez-Perez, X. Chu, and İ. Guvenc. On the expanded region of picocells in heterogeneous networks. *IEEE J.of Selected Topics in Sig. Processing*, 6(3):281–294, Jun. 2012.

[83] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji. Cell association and interference coordination in heterogeneous LTE-A cellular networks. *IEEE J.on Selected Areas in Commun.*, 28(9):1479–1489, 2010.

[84] D. Malak, H. S. Dhillon, and J. G. Andrews. Optimizing data aggregation for uplink machine-to-machine communication networks. *IEEE Trans. Commun.*, 64(3):1274–1290, Mar. 2016.

[85] Francisco J Martin-Vega, Gerardo Gomez, Mari Carmen Aguayo-Torres, and Marco Di Renzo. Analytical modeling of interference aware power control for the uplink of heterogeneous cellular networks. *IEEE Trans. Wireless Commun.*, 15(10):6742–6757, Oct. 2016.

[86] E. Matskani, N. D. Sidiropoulos, Z. Luo, and L. Tassiulas. Convex approximation techniques for joint multiuser downlink beamforming and admission control. *IEEE Trans. on Wireless Commun.*, 7(7):2682–2693, Jul. 2008.

[87] G. Miao, A. Azari, and T. Hwang. $E^2$ -MAC: Energy efficient medium access for massive M2M communications. *IEEE Trans. Commun.*, 64(11):4720–4735, Nov. 2016.

[88] J. Mišić and V. B. Mišić. Efficiency of power ramping during random access in LTE. *IEEE Trans. on Veh. Tech.*, 67(2):1698–1712, 2018.

[89] J. Mišić, V. B. Mišić, and N. Khan. Sharing it my way: Efficient M2M access in LTE/LTE-A networks. *IEEE Trans. on Veh. Tech.*, 66(1):696–709, 2017.

[90] H. G. Moussa and W. Zhuang. RACH performance analysis for large-scale cellular IoT applications. *IEEE Internet Things J.*, 6(2):3364–3372, Apr. 2019.

[91] H. G. Moussa and W. Zhuang. Access point association in uplink two-hop cellular IoT networks with data aggregators. *IEEE Internet Things J.*, pages 1–1, Mar. 2020.

[92] H. G. Moussa and W. Zhuang. Energy- and delay-aware two-hop NOMA-enabled massive cellular IoT communications. *IEEE Internet Things J.*, 7(1):558–569, Jan.. 2020.

[93] K. S. NataraJan.. A hybrid medium access control protocol for wireless LANs. In *Proc. IEEE Inter. Conf. on Selected Topics in Wireless Commun.*, pages 134–137, Jun. 1992.

[94] D. W. K. Ng, Y. Wu, and R. Schober. Power efficient resource allocation for full-duplex radio distributed antenna networks. *IEEE Trans. on Wireless Commun.*, 15(4):2896–2911, Apr. 2016.

[95] T. D. Novlan, H. S. Dhillon, and J. G. Andrews. Analytical modeling of uplink cellular networks. *IEEE Trans. Wireless Commun.*, 12(6):2669–2679, Jun. 2013.

[96] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, 2009.

[97] H. Pervaiz, L. Musavian, and Q. Ni. Joint user association and energy-efficient resource allocation with minimum-rate constraints in two-tier HetNets. In *Proc. IEEE PIMRC*, pages 1634–1639, Sep. 2013.

[98] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Difference of convex functions programming for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2519–2527, 2014.

[99] M. Polese, M. Centenaro, A. Zanella, and M. Zorzi. M2M massive access in LTE: RACH performance evaluation in a smart city scenario. In *Proc. IEEE ICC*, pages 1–6, May 2016.

[100] Tran Dinh Quoc and Moritz Diehl. Sequential convex programming methods for solving nonlinear optimization problems with dc constraints. *arXiv preprint arXiv:1107.5841*, 2011.

[101] A. Rico-Alvarino, M. Vajapeyam, H. Xu, X. Wang, Y. Blankenship, J. Bergman, T. Tirronen, and E. Yavuz. An overview of 3GPP enhancements on machine to machine communications. *IEEE Commun. Mag.*, 54(6):14–21, Jun. 2016.

[102] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor. A college admissions game for uplink user association in wireless small cell networks. In *Proc. IEEE INFOCOM*, pages 1096–1104, Apr. 2014.

[103] S. Sadr and R. S. Adve. Tier association probability and spectrum partitioning for maximum rate coverage in multi-tier heterogeneous networks. *IEEE Commun. Letters*, 18(10):1791–1794, Oct. 2014.

[104] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi. Non-orthogonal multiple access (NOMA) for cellular future radio access. In *Proc. IEEE VTC*, pages 1–5, Jun. 2013.

[105] T. Salam, W. U. Rehman, and X. Tao. Cooperative data aggregation and dynamic resource allocation for massive machine type communication. *IEEE Access*, 6:4145 – 4158, 2018.

[106] N. Saquib, E. Hossain, L. B. Le, and D. I. Kim. Interference management in OFDMA femtocell networks: Issues and approaches. *IEEE Wireless Commun.*, 19(3):86–95, Jun. 2012.

[107] K. Shen and W. Yu. Distributed pricing-based user association for downlink heterogeneous cellular networks. *IEEE J.on Selected Areas in Commun.*, 32(6):1100–1113, Jun. 2014.

[108] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson. Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations. *IEEE Commun. Mag.*, 55(9):55–61, Sep. 2017.

[109] M. Shirvanimoghaddam and S. Johnson. Multiple access technologies for cellular M2M communications: An overview. *arXiv preprint arXiv:1611.05548*, 2016.

[110] S. Singh, H. S. Dhillon, and J. G. Andrews. Offloading in heterogeneous networks: Modeling, analysis, and design insights. *IEEE Trans. on Wireless Commun.*, 12(5):2484–2497, 2013.

[111] S. Singh, X. Zhang, and J. G. Andrews. Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in HetNets. *IEEE Trans. Wireless Commun.*, 14(10):5360–5373, Oct. 2015.

[112] K. Smiljkovikj, P. Popovski, and L. Gavrilovska. Analysis of the decoupled access for downlink and uplink in wireless heterogeneous networks. *IEEE Wireless Commun. Letters*, 4(2):173–176, Apr. 2015.

[113] H. Tabassum, E. Hossain, and J. Hossain. Modeling and analysis of uplink non-orthogonal multiple access (NOMA) in large-scale cellular networks using poisson cluster processes. *IEEE Trans. Commun.*, 65(8):3555–3570, Aug. 2017.

[114] T. Takeda and K. Higuchi. Enhanced user fairness using non-orthogonal access with SIC in cellular uplink. In *Proc. IEEE VTC*, pages 1–5, Sep. 2011.

[115] P. D. Tao et al. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of operations research*, 133(1-4):23–46, 2005.

[116] H. Tian, W. Xie, X. Gan, and Y. Xu. Hybrid user association for maximising energy efficiency in heterogeneous networks with human-to-human/machine-to-machine coexistence. *IET Commun.*, 10(9):1035–1043, Jun. 2016.

[117] R. Trestian, O. Ormond, and G. Muntean. Game theory-based network selection: Solutions and challenges. *IEEE Commun. Surveys & Tuts.*, 14(4):1212–1231, Feb. 2012.

[118] M. Vilgelm, H. M. Gürsu, W. Kellerer, and M. Reisslein. LATMAPA: Load-adaptive throughput- maximizing preamble allocation for prioritization in 5G random access. *IEEE Access*, 5:1103–1116, Jan... 2017.

[119] P. Wang, J. Xiao, and L. P. Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems. *IEEE Veh. Tech. Mag.*, 1(3):4–11, 2006.

[120] D. T. Wiriaatmadja and K. W. Choi. Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks. *IEEE Trans. Wireless Commun.*, 14(1):33–46, Jan... 2015.

[121] X. Yang, A. Fapojuwo, and E. Egbogah. Performance analysis and parameter optimization of random access backoff algorithm in LTE. In *Proc. IEEE VTC*, pages 1–5, Sep. 2012.

[122] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir. A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems. *IEEE Trans. on Wireless Commun.*, 15(11):7244–7257, Nov. 2016.

[123] Z. Yang, W. Xu, H. Xu, J. Shi, and M. Chen. Energy efficient non-orthogonal multiple access for machine-to-machine communications. *IEEE Commun. Letters*, 21(4):817–820, Apr. 2017.

[124] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu. Multi-user shared access for Internet of Things. In *Proc. IEEE VTC*, pages 1–5, May 2016.

[125] R. Zhang, M. Wang, X. Shen, and L. L. Xie. Probabilistic analysis on QoS provisioning for internet of things in LTE-A heterogeneous networks with partial spectrum usage. *IEEE Internet Things J.*, 3(3):354–365, Jun. 2016.

[126] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler. Radio resource allocation in LTE-advanced cellular networks with M2M communications. *IEEE Commun. Mag.*, 50(7):184–192, Jul. 2012.

# Appendix A

# RACH performance Analysis

## A.1 Association success probability

Device$_0$ located at $d_{00}^n$ successfully associates with the BS$_0$ located at the origin if its transmitted preamble is received with SINR above threshold $\tau$. The probability of successful association is given by

$$\begin{aligned}
\Theta(\Xi_{00}^n) &= P\left(\frac{q_d||d_{00}^n||^{(\epsilon-1)\alpha} g_0}{I_{in}+I_{out}+\sigma^2} > \tau\right) \\
&= E_{||d_{00}^n||,I_{in},I_{out}}\left[P(g_0 > \frac{\tau}{q_d}(I_{in}+I_{out}+\sigma^2)||d_{00}^n||^{(1-\epsilon)\alpha})\right] \\
&\stackrel{a}{=} E_{||d_{00}^n||,I_{in},I_{out}}\left[\exp\{-\frac{\tau}{q_d}(I_{in}+I_{out}+\sigma^2)||d_{00}^n||^{(1-\epsilon)\alpha}\}\right] \\
&\stackrel{b}{=} E_{||d_{00}^n||,I_{in},I_{out}}\Big[\exp\{-\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha}\sigma^2\} \\
&\qquad \times \exp\{-\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha} I_{in}\}\exp\{-\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha} I_{out}\}\Big] \\
&\stackrel{c}{=} E_{||d_{00}^n||}\left[\exp\{-\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha}\sigma^2\}\right]\mathscr{L}_{I_{in}}\{\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha}\} \\
&\qquad \times \mathscr{L}_{I_{out}}\{\frac{\tau}{q_d}||d_{00}^n||^{(1-\epsilon)\alpha}\}
\end{aligned} \tag{A.1}$$

where $E[.]$ denotes the expectation operator, (a) follows from the fact that the channel gain $g$ is exponentially distributed with unity mean, (b) is the expectation with respect to $||d_{00}^n||$, $I_{in}$ and $I_{out}$ which are random due to the randomness in the link lengths between devices and their serving BSs, while the $\mathscr{L}_{I_{in}}\{.\}$ and $\mathscr{L}_{I_{out}}\{.\}$ terms in (c) denote the Laplace transforms of the intra-cell and inter-cell interference respectively.

## A.2 Average number of active devices per Voronoi cell

$$\bar{M}_0 \stackrel{\triangle}{=} E[M_0] = \sum_{m=0}^{\infty} mP(M_0 = m)$$

$$
\begin{aligned}
&= \frac{\lambda_b^c}{\Gamma(c)(\lambda_b+\lambda_d)^c} \sum_{m=0}^{\infty} \frac{m\lambda_d^m \Gamma(m+c)}{\Gamma(m+1)(\lambda_b+\lambda_d)^m} \\
&\overset{a}{=} \gamma \int_0^{\infty} \sum_{m=0}^{\infty} \frac{m\zeta^m t^{m+c-1}}{m!} e^{-t} dt \\
&\overset{b}{=} \gamma \int_0^{\infty} \zeta t^c e^{(\zeta-1)t} dt = \frac{\gamma\zeta\Gamma(c+1)}{(1-\zeta)^{(c+1)}}
\end{aligned} \tag{A.2}
$$

where (a) follows from the definition of Gamma function ($\Gamma(a+c) = \int_0^{\infty} t^{a+c-1}e^{-t}$) which is true for $a < 1$, and (b) follows from the summation relation ($\sum_{a=0}^{\infty} \zeta^a/a! = e^a$).

## A.3 Laplace transform of intra-cell interference

Derivations of the intra-cell interference on the $n^{th}$ preamble are as follows:

$$
\begin{aligned}
\mathscr{L}_{I_{in}}\{s\} &= E_{I_{in}}\left[\exp\{-sq_d \sum_{d_{0i}^n \in \Psi_{n0}/\{d_{00}^n\}} (R_{0i}^n)^{(\epsilon-1)\alpha} g_i\}\right] \\
&\overset{a}{=} E_{R_{0i}^n, g_i}\left[\prod_{d_{0i}^n \in \Psi_{n0}/\{d_{00}^n\}} \exp\{-sq_d (R_{0i}^n)^{(\epsilon-1)\alpha} g_i\}\right] \\
&= E_{R_{0i}^n}\left[\prod_{d_{0i}^n \in \Psi_{n0}/\{d_{00}^n\}} E_{g_i}[\exp\{-sq_d (R_{0i}^n)^{(\epsilon-1)\alpha} g_i\}]\right] \\
&\overset{b}{=} E_{R_{0i}^n}\left[\prod_{d_{0i}^n \in \Psi_{n0}/\{d_{00}^n\}} \left(\frac{1}{1+sq_d (R_{0i}^n)^{(\epsilon-1)\alpha}}\right)\right] \\
&\overset{c}{=} \left(\int_0^{\infty} \frac{f_{R_{0i}^n}(r)}{1+sr^{(\epsilon-1)\alpha}} dr\right)^{\bar{M}} \\
&\overset{d}{=} \left(\int_0^{\infty} \frac{f_{R^n}(r)}{1+sq_d r^{(\epsilon-1)\alpha}} dr\right)^{\bar{M}}
\end{aligned} \tag{A.3}
$$

where (a) follows from the independence of $g_i$ (for different values of $i$), (b) follows from $g_i$ being exponentially distributed with unity mean and its independence of the propagation distance, and (c) follows from the i.i.d nature of the distances from the devices to $BS_0$. Since $R_{0i}^n$ are i.i.d for different values of $i$ and are independent of the BS location, in (d), the subscript $\{0i\}$ is omitted such that the PDF is denoted by $f_{R^n}(r)$ given by (3.4).

## A.4 Laplace transform of inter-cell interference

Derivations of the inter-cell interference on the $n^{th}$ preamble are as follows:

$$
\mathscr{L}_{I_{out}}\{s\} = E_{g_i, s, ||d_{ji}^n - b_j||, ||d_{ji}^n||}\left[\exp\{-sq_d \sum_{j\neq 0} \sum_{d_{ji}^n \in \Psi_{nj}} ||d_{ji}^n||^{-\alpha} ||d_{ji}^n - b_j||^{\epsilon\alpha} g_i\}\right]
$$

$$
\begin{aligned}
&= E_{X_{ji}^n, Y_{ji}^n, g_i}\left[\prod_{j=1}\prod_{d_{ji}^n \in \Psi_{nj}} \exp\{-sq_d (Y_{ji}^n)^{\epsilon\alpha} (X_{ji}^n)^{-\alpha} g_i\}\right] \\
&= E_{X_{ji}^n, Y_{ji}^n}\left[\prod_{j=1}\prod_{d_{ji}^n \in \Psi_{nj}} E_{g_i}[\exp\{-sq_d (Y_{ji}^n)^{\epsilon\alpha} (X_{ji}^n)^{-\alpha} g_i\}]\right] \\
&\overset{a}{=} E_{X_{ji}^n}\left[\prod_{j=1}\left(E_{Y_{ji}^n}\left[\frac{1}{1+sq_d (Y_{ji}^n)^{\epsilon\alpha} (X_{ji}^n)^{-\alpha}}\right]\right)^{\bar{M}}\right] \\
&\overset{b}{=} \exp\left(-2\pi\lambda_n \int_{x>0} (1-\exp\left(-\lambda_n \pi x^2\right))\left[\left(1-\left(E_{Y_{ji}^n}\left[\frac{1}{1+sq_d (Y_{ji}^n)^{\epsilon\alpha} (X_{ji}^n)^{-\alpha}}\right]\right)^{\bar{M}}\right)\right] x dx\right) \\
&\overset{c}{=} \exp\left(-2\pi\lambda_n \int_{x>0} (1-\exp\left(-\lambda_n \pi x^2\right))\left[1-\left(\int_0^\infty \frac{f_{Y^n}(y)}{1+sq_d y^{\epsilon\alpha} x^{-\alpha}} dy\right)^{\bar{M}}\right] x dx\right) \\
&\overset{d}{=} \exp\left(-2\pi\lambda_n \int_{x>0} \left[1-\left(\int_0^\infty \frac{f_{Y^n}(y)}{1+sq_d y^{\epsilon\alpha} x^{-\alpha}} dy\right)^{\bar{M}}\right] x dx\right)
\end{aligned} \tag{A.4}
$$

where (a) follows from the i.i.d. exponentially distributed channel gain $g_i$ with unity mean (for different $i$ values) and its independence from the propagation distance, $Y_{ji}^n$, and the fact that $Y_{ji}^n$ is i.i.d for different values of $i$: (b) follows from $X_{ji}^n$ being i.i.d for different values of $j$ and from the definition of the probability generating functionl (PGFL) of the PPP [51]. In (c), the subscripts are omitted in $f_{Y^n}(y)$ which is the i.i.d PDF of $Y_{ji}^n$ (for different values of $i$ and $j$) given by (A.5) shown below. In (d), we assume that the locations of inter-cell interfering devices follow a homogeneous PPP for simplicity.

$$
f_{Y_{ji}^n}(y) = 2\pi\lambda_b y \exp\left\{(-\pi\lambda_b y^2)\right\}. \tag{A.5}
$$

# Appendix B

# NOMA-Enabled Two-hop Cellular Network

## B.1 NOMA sub-channels scheduling probability

We define scheduling probability as the probability of having $l$ scheduled devices on the $n^{th}$ NOMA sub-channel. This probability depends on the number of active devices in a cluster and the number of available resources. Accordingly, the conditional PMF of $S^n$ given $\mathcal{M}_a = m$ is

$$P(S^n = s|\mathcal{M}_a = m) = \begin{cases} 1 - \frac{m}{\mathcal{N}}, & s = 0,\ m < \mathcal{N} \\ \frac{m}{\mathcal{N}}, & s = 1,\ m < \mathcal{N} \\ 2 - \frac{m}{\mathcal{N}}, & s = 1,\ \mathcal{N} \leq m < M_{max} \\ \frac{m}{\mathcal{N}} - 1, & s = 2,\ \mathcal{N} \leq m < M_{max} \\ 1, & s = 2,\ m \geq M_{max} \\ 0, & \text{otherwise.} \end{cases}$$

To find the unconditional PMF, $P(S^n = s) = P(S^n = s|\mathcal{M}_0 = m)P(\mathcal{M}_0 = m)$, we use the PDF of $\mathcal{M}_0$ given in (1) to yield the following

$$P(S^n = s) = \begin{cases} \sum_{m=0}^{\mathcal{N}-1} (1 - \frac{m}{\mathcal{N}})P(\mathcal{M}_0 = m), & s = 0,\ m < \mathcal{N} \\ \sum_{m=0}^{\mathcal{N}-1} (\frac{m}{\mathcal{N}})P(\mathcal{M}_0 = m), & s = 1,\ m < \mathcal{N} \\ \sum_{m=\mathcal{N}}^{M_{max}-1} (2 - \frac{m}{\mathcal{N}})P(\mathcal{M}_0 = m), & s = 1,\ \mathcal{N} \leq m < M_{max} \\ \sum_{m=\mathcal{N}}^{M_{max}-1} (\frac{m}{\mathcal{N}} - 1)P(\mathcal{M}_0 = m), & s = 2,\ \mathcal{N} \leq m < M_{max} \\ \sum_{m=M_{max}}^{\infty} P(\mathcal{M}_0 = m), & s = 2,\ m \geq M_{max} \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \sum_{m=0}^{\mathcal{N}-1} (1 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 0 \\ \sum_{m=\mathcal{N}}^{M_{max}-1} (2 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m) + \sum_{m=0}^{\mathcal{N}-1} (\frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 1 \\ \sum_{m=M_{max}}^{\infty} P(\mathcal{M}_a = m) + \sum_{m=\mathcal{N}}^{M_{max}-1} (\frac{m}{\mathcal{N}} - 1) P(\mathcal{M}_a = m), & s = 2 \\ 0, & s > 2. \end{cases}$$

## B.2 Device-DA probability of successful transmission

By definition, the device-DA transmission success probability, $\Theta(\Xi_{0j}^{ns})$, for different scenarios is given by

$$\begin{aligned}
\Theta(\Xi_{0h}^{n1}) &= P\left\{\Xi_{0h}^{n1} > \tau_{ah}\right\} = P\left\{\frac{g_{0h}^{n0}}{I_{2h} + I_{2l}} > \tau_{ah}\right\} \\
&= E_{I_{2h}, I_{2l}}\left[\exp\left\{-\tau_{ah}(I_{2h} + I_{2l})\right\}\right] \\
&= \mathscr{L}_{I_{2h}}\left\{\tau_{ah}\right\} \mathscr{L}_{I_{2l}}\left\{\tau_{ah}\right\} \\
\Theta(\Xi_{0h}^{n2}) &= P\left\{\frac{g_{0h}^{n0}}{g_{0l}^{n0} + I_{2h} + I_{2l}} > \tau_{ah}\right\} \\
&= E_{g_{0l}^{n0}}\left[\exp\{-\tau_{ah} g_{0l}^{n0}\}\right] \mathscr{L}_{I_{2h}}\left\{\tau_{ah}\right\} \mathscr{L}_{I_{2l}}\left\{\tau_{ah}\right\} \\
\Theta(\Xi_{0l}^{n2}) &= P\left\{\frac{g_{0l}^{n0}}{I_{2h} + I_{2l}} > \tau_{al} \cap \frac{g_{0h}^{n0}}{g_{0l}^{n0} + I_{2h} + I_{2l}} > \tau_{ah}\right\} \\
&\overset{a}{=} \int_{\tau_{al}(I_{2h}+I_{2l})}^{\infty} e^{-g_{02}^{n0}} \int_{\tau_{ah}(g_{0l}^{n0}+I_{2h}+I_{2l})}^{\infty} e^{-g_{0h}^{n0}} \, dg_{0h}^{n0} \, dg_{0l}^{n0} \\
&= \frac{1}{1 + \tau_{ah}} E_{I_{2h}, I_{2l}}\left[\exp\left\{-(\tau_{ah} + \tau_{al}(1 + \tau_{ah}))(I_{2h} + I_{2l})\right\}\right] \\
&\overset{b}{=} \frac{1}{1 + \tau_{ah}} \mathscr{L}_{I_{2h}}\left\{A\right\} \mathscr{L}_{I_{2l}}\left\{A\right\}
\end{aligned}$$

where (a) follows from $g_{0h}^{n0}$ and $g_{0l}^{n0}$ being i.i.d exponentially distributed RVs with unity mean, and in (b) $A = \tau_{ah} + \tau_{al}(1 + \tau_{ah})$.

For simplicity of notation, let $Y_{ij}^{n} = ||x_{ij}^{n} - a_k||$ and $R_{i0}^{n} = ||x_{ij}^{n}||$ denote the distances from the $j$ ranked $i^{th}$ interfering device on the $n^{th}$ sub-channel, located at $x_{ij}^{n}$, to its serving DA and to the origin respectively. Since device locations follow a PPP with node density $\lambda_d$ and devices are randomly chosen and scheduled on the available sub-channels, $Y_{ij}^{n}$ is an i.i.d RV for different values of $i$ and $j$, and follows a Rayleigh distribution with parameter $\lambda_d$ [51]. Similarly, $R_{i0}^{n}$ follows a Rayleigh distribution with parameter $\lambda_d^j$, for $j \in \{h, l\}$. Accordingly, the Laplace

transforms, $\mathscr{L}_{I_{2j}}\{\kappa\}$, for $\kappa \in \{\tau_{ah}, A\}$ and $j \in \{h, l\}$, are given by

$$\begin{aligned}
\mathscr{L}_{I_{2j}}\{\kappa\} &= E_{I_{2j}}\left[e^{\left\{-\kappa \sum_{x_{ij}^n \in \Psi_n^j} \mathbb{1}(Y_{ij}^n < R_{ij}^n)(Y_{ij}^n)^\alpha (R_{ij}^n)^{-\alpha} g_{ij}^{n0}\right\}}\right] \\
&\overset{a}{=} E_{\Psi_n^j}\left[\prod_{x_{ij}^n \in \Psi_n^j} E_{Y_{ij}^n, g_{ij}^{n0}}\left[e^{\left\{-\kappa \mathbb{1}\left(Y_{ij}^n < R_{ij}^n\right)(Y_{ij}^n)^\alpha (R_{ij}^n)^{-\alpha} g_{ij}^{n0}\right\}}\right]\right] \\
&\overset{b}{=} \exp\left\{-2\pi\lambda_d^j E_{Y_{ij}^n, g_{ij}^{n0}}\left[\int_y^\infty (1 - e^{-\kappa (Y_{ij}^n)^\alpha r^{-\alpha} g_{ij}^{n0}}) r \, dr\right]\right\} \\
&\overset{c}{=} \exp\left\{-2\pi\lambda_d^j E_{Y_{ij}^n}\left[\int_y^\infty \left(\frac{1}{(\kappa (Y_{ij}^n)^\alpha)^{-1} r^\alpha + 1}\right) r \, dr\right]\right\} \\
&\overset{d}{=} \exp\left\{-\pi\lambda_d^j \kappa^{\frac{2}{\alpha}} E_{Y_{ij}^n}[((Y_{ij}^n)^\alpha)^{\frac{2}{\alpha}}] \int_{(\kappa)^{\frac{-2}{\alpha}}}^\infty \frac{1}{z^{\frac{\alpha}{2}} + 1} \, dz\right\}
\end{aligned}$$

where $E_X[\cdot]$ is the expectation with respect to the variable X, (a) follows from the independence between $\Psi_n^j$ and $g_{ij}^{n0}$, (b) follows from the probability generation functional (PGFL) of the PPP [51], (c) follows from the Laplace transform of $g_{ij}^{n0}$ and the fact that $g_{ij}^{n0}$ are i.i.d exponential random variables with unity mean for different values of $i$ and $j$. As for (d), it is obtained with introducing variables $z_{ij} = r^2/(\kappa (Y_{ij}^n)^\alpha)^{2/\alpha}$. Using the PDF of the distance between an arbitrary device and its serving DA, the term $E_{Y_{ij}^n}[((Y_{ij}^n)^\alpha)^{\frac{2}{\alpha}}]$ is given by

$$\begin{aligned}
E_{Y_{ij}^n}[((Y_{ij}^n)^\alpha)^{\frac{2}{\alpha}}] &= \int_0^\infty (y^\alpha)^{\frac{2}{\alpha}} f_Y(y) \, dr \\
&= \int_0^\infty y^2 2\pi\lambda_d y e^{-\pi\lambda_d y^2} \, dy \\
&= \frac{1}{\pi\lambda_d}.
\end{aligned}$$

Accordingly, at $\alpha = 4$, the Laplace transforms, $\mathscr{L}_{I_{2j}}\{\kappa\}$, for $\kappa \in \{\tau_{ah}, A\}$ and $j \in \{h, l\}$ simplify to

$$\mathscr{L}_{I_{2j}}\{\kappa\} = \exp\left\{-\frac{\sqrt{\kappa}\lambda_d^j}{\lambda_d}\left(\frac{\pi}{2} - \arctan\left((\kappa)^{-1/2}\right)\right)\right\}.$$

## B.3 Average achievable bit rate by a device in coverage

A device can transmit to its serving DA if and only if it is in coverage. Accordingly, the achievable bit rate is conditional on being in coverage. The average achievable bit rate by a device, when in coverage, can be derived using the Shannon's channel capacity formula, given by

$$\mathcal{F}_d^{sjn} = \begin{cases} E\left[\omega_n \log_2\left(1+\Xi_{0h}^{n1}\right)\right], & s=1 \;\&\; j=h \\ E\left[\omega_n \log_2\left(1+\Xi_{0h}^{n2}\right)\right], & s=2 \;\&\; j=h \\ E\left[\omega_n \log_2\left(1+\Xi_{0l}^{n2}\right)\right], & s=2 \;\&\; j=l \end{cases}$$

where

$$\begin{aligned}
\mathcal{F}_d^{1h} &= \int_{t>0} P\left(\omega_n \log_2\left(1+\Xi_{0h}^{n1}\right) > t\right) dt \\
&= \int_{t>0} P\left(\Xi_{0h}^{n1} > 2^{\frac{t}{\omega_n}} - 1\right) dt \\
&\overset{a}{=} \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \\
\mathcal{F}_d^{2h} &= \int_{t>0} P\left(\omega_n \log_2\left(1+\Xi_{0h}^{n2}\right) > t\right) dt \\
&= \int_{t>0} P\left(\Xi_{0h}^{n2} > 2^{\frac{t}{\omega_n}} - 1\right) dt \\
&\overset{b}{=} \int_{t>0} E_{g_{0l}^{n0}}\left[\exp\{-\mathcal{B} g_{0l}^{n0}\}\right] \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \\
\mathcal{F}_d^{1h} &= \int_{t>0} P\left(\omega_n \log_2\left(1+\Xi_{0l}^{n2}\right) > t\right) dt \\
&= \int_{t>0} P\left(\Xi_{0l}^{n2} > 2^{\frac{t}{\omega_n}} - 1\right) dt \\
&\overset{c}{=} \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt
\end{aligned}$$

where $\mathcal{B} = 2^{\frac{t}{\omega_n}} - 1$, and (a), (b), and (c) follow from that the probability terms resemble those of the coverage probabilities given in (40). We use the expectation with respect to interference terms to reach the Laplace transform terms.

# Appendix C

# Optimized Data Aggregation

## C.1 Average achievable data rate of a DA

Using Shannon's channel capacity formula, the achievable data rate by a DA is given as

$$\begin{aligned}
\bar{R}_a &= E_{X_j,Y_j,g^l_{a_j,0},g^l_{a_0,0}}\left[\omega_u \log_2\left(1+\frac{q_o g^l_{a_0,0}}{\sum\limits_{a^l_j\in\Phi^l_a/a^l_0} q_o Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}}\right)\right] \\
&\overset{a}{=} \int_{t>0} P\left(\omega_u \log_2\left(1+\frac{q_o g^l_{a_0,0}}{\sum\limits_{a^l_j\in\Phi^l_a/a^l_0} q_o Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}}\right) > t\right) \mathrm{d}t \\
&= \int_{t>0} P\left(g^l_{a_0,0} > \left(2^{\frac{t}{\omega_u}}-1\right) \sum_{a^l_j\in\Phi^l_a/a^l_0} Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}\right) \mathrm{d}t \\
&\overset{b}{=} \int_{t>0} E_{X_j,Y_j,g^l_{a_j,0}}\left[\exp\left\{-\tau \sum_{a^l_j\in\Phi^l_a/a^l_0} Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}\right\}\right] \mathrm{d}t \\
&= \int_{t>0} \mathscr{L}_{X_j,Y_j,g^l_{a_j,0}}\{\tau\} \mathrm{d}t
\end{aligned}$$

where $E_X[\cdot]$ is the expectation with respect to $X$, and in (a), we use the definition $E[X] = \int_{t>0} P(X>t)\, dt$ to compute the expectation of the logarithmic term; in (b), $\tau = 2^{\frac{t}{\omega_u}} - 1$, and the exponential function follows from the exponentially distributed channel power gain $g^l_{a_0,0}$ with unity mean. Further, $\mathscr{L}_X\{\cdot\}$ denotes Laplace transform with respect to $X$, given by

$$\begin{aligned}
\mathscr{L}_{X_j,Y_j,g^l_{a_j,0}}\{\tau\} &= E_{X_j,Y_j,g^l_{a_j,0}}\left[\exp\left\{-\tau \sum_{a^l_j\in\Phi^l_a/a^l_0} Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}\right\}\right] \\
&\overset{a}{=} E_{X_j}\left[\prod_{a^l_j\in\Phi^l_a/a^l_0} E_{Y_j,g^l_{a_j,0}}\left[e^{\left\{-\tau Y_j^{\alpha} X_j^{-\alpha} g^l_{a_j,0}\right\}}\right]\right]
\end{aligned}$$

$$\stackrel{b}{=} \exp\left\{-2\pi\lambda_b E_{Y_j, g^l_{a_j,0}}\left[\int_y^\infty (1 - e^{-\tau Y_j^\alpha x^{-\alpha} g^l_{a_j,0}}) x \, \mathrm{d}x\right]\right\}$$
$$\stackrel{c}{=} \exp\left\{-2\pi\lambda_b E_{Y_j}\left[\int_y^\infty \left(\frac{1}{\tau Y_j^\alpha x^{-\alpha} g^l_{a_j,0} + 1}\right) x \, \mathrm{d}x\right]\right\}$$
$$\stackrel{d}{=} \exp\left\{-\pi\lambda_b \tau^{\frac{2}{\alpha}} E_{Y_j}[((Y_j)^\alpha)^{\frac{2}{\alpha}}] \int_{(\tau)^{\frac{-2}{\alpha}}}^\infty \frac{1}{z^{\frac{\alpha}{2}} + 1} \, \mathrm{d}z\right\}$$

where (a) follows from the independence between $\mathbf{\Phi}^l_a$ and $g^l_{a_j,0}$, (b) follows from the probability generation functional of the PPP [51], (c) follows from the exponentially distributed random variable $g^l_{a_j,0}$ with unity mean and the i.i.d $g^{n0}_{ij}$ for different values of $j$. As for (d), it is obtained with introducing variable $z_j = r^2/(\tau(Y_j)^\alpha)^{2/\alpha}$. Using the PDF of the distance between an arbitrary device and its serving DA, the term $E_{Y_j}[((Y_j)^\alpha)^{\frac{2}{\alpha}}]$ is given by

$$\begin{aligned} E_{Y_j}[((Y_j)^\alpha)^{\frac{2}{\alpha}}] &= \int_0^\infty (y^\alpha)^{\frac{2}{\alpha}} f_Y(y) \, \mathrm{d}r \\ &= \int_0^\infty y^2 2\pi\lambda_a y e^{-\pi\lambda_a y^2} \, dy = \frac{1}{\pi\lambda_a}. \end{aligned}$$

Accordingly, at $\alpha = 4$, the Laplace transforms, $\mathscr{L}_{X_j, Y_j, g^l_{a_j,0}}\{\tau\}$, simplifies to

$$\mathscr{L}_{X_j, Y_j, g^l_{a_j,0}}\{\tau\} = \exp\left\{-\frac{\sqrt{\tau}\lambda_b}{\lambda_a}\left(\frac{\pi}{2} - \arctan\left((\tau)^{-1/2}\right)\right)\right\}.$$

## C.2 Proof of equivalency of problem P1 and problem P2

Let $(R^*, S^*, X^*, Y^*, P^*)$ be the optimal solution to problem **P1**. Further, let $(\widetilde{R}, \widetilde{S}, \widetilde{X}, \widetilde{Y}, \widetilde{P})$ be an optimal solution to problem **P2**. Since the set of constraints of problem **P1** and problem **P2** are the same, it follows that $(R^*, S^*, X^*, Y^*, P^*)$ is also a feasible solution of problem **P2**. In what follows, we prove that $(R^*, S^*, X^*, Y^*, P^*)$ is not only another feasible solution of problem **P2** but in fact the optimal solution of it. First, for simplicity of notations, define the following

$$a^* = \left(\sum_{b_j \in \Psi_b} \sum_{d_i \in \Psi_d} x^*_{ij} + \sum_{a_h \in \Psi_a} \sum_{d_i \in \Psi_d} y^*_{ih}\right) \tag{C.1}$$
$$b^* = \sum_{d_i \in \Psi_d} (x^*_{ij} + y^*_{ih}) q_{d_i} \tag{C.2}$$
$$\widetilde{a} = \left(\sum_{b_j \in \Psi_b} \sum_{d_i \in \Psi_d} \widetilde{x}_{ij} + \sum_{a_h \in \Psi_a} \sum_{d_i \in \Psi_d} \widetilde{y}_{ih}\right) \tag{C.3}$$
$$\widetilde{b} = \sum_{d_i \in \Psi_d} (\widetilde{x}_{ij} + \widetilde{y}_{ih}) q_{d_i} \tag{C.4}$$

Since $x^*_{ij}$, $y^*_{ih}$, $\widetilde{x}_{ij}$ and $\widetilde{y}_{ih}$ are all positive integers that assumes a maximum value of 1, define $\Delta A$ as

$$(a^* - \widetilde{a}) = \Delta A \geq 1 \tag{C.5}$$

Accordingly, using the preceding definitions and the formulation of problem **P1** and problem **P2**, we have the following

$$\epsilon \Delta B - (1-\epsilon)\Delta A \leq \frac{\Delta B}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}} - \frac{\Delta A \sum_{q_{d_i} \in \Psi_d} Q_{d_i,max}}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}} \tag{C.6}$$

$$\leq \frac{\Delta B}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}} - \frac{\sum_{q_{d_i} \in \Psi_d} Q_{d_i,max}}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}} = \sigma \tag{C.7}$$

where $\Delta B = b^* - \widetilde{b}$, (C.6) follows from the definition $0 \leq \epsilon \leq \frac{1}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}}$, and (C.7) follows from the fact that the minimum value $\Delta A$ can assume, based on (C.5), is 1. Furthermore, notice that, since $-\sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max} \leq \Delta B \leq \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}$, then

$$\frac{-2 \sum_{q_{d_i} \in \Psi_d} Q_{d_i,max}}{1 + \sum\limits_{q_{d_i} \in \Psi_d} Q_{d_i,max}} \leq \sigma \leq 0 \tag{C.8}$$

As we assume that $(R^*, S^*, X^*, Y^*, P^*)$ and $(\widetilde{R}, \widetilde{S}, \widetilde{X}, \widetilde{Y}, \widetilde{P})$ are both feasible solutions for problem **P2**, than they should result in the optimal objective function. Thus, we can conclude that

$$\epsilon \Delta B - (1-\epsilon)\Delta A \leq 0 \tag{C.9}$$

$$\epsilon(b^* - \widetilde{b}) - (1-\epsilon)(a^* - \widetilde{a}) \leq 0 \tag{C.10}$$

$$\epsilon b^* - (1-\epsilon)a^* \leq \epsilon \widetilde{b} - (1-\epsilon)\widetilde{a} \tag{C.11}$$

Accordingly, as $(R^*, S^*, X^*, Y^*, P^*)$ results in a smaller objective function in problem **P2** compared to $(\widetilde{R}, \widetilde{S}, \widetilde{X}, \widetilde{Y}, \widetilde{P})$ do, hence, we can conclude that $(R^*, S^*, X^*, Y^*, P^*)$ is in fact the optimal solution of both problem **P1** as well as problem **P2**. ■

## C.3 Proof of Proposition 1

Notice that, the following proof is applicable to both problem **P4-1** and problem **P4-2**. However, we do the proof only for problem **P4-1**.

Similar to [94], we start the proof by using the *abstract Lagrangian duality*. First, for ease of notation, let $\Pi(S, \mathbf{Q}_b) = \sum\limits_{d_i \in C_{1,h}}^{C_{\mathcal{B},h}} \sum\limits_{k=1}^{K} s^k_{d_i, C_{j,h}} q^k_{d_i, C_{j,h}}$. Accordingly, by applying the Lagrangian to combine constraint $(C4 - 1.3.2)$ into the objective function of problem **P4-1**, the optimization problem in (5.7) can be written as

$$\underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \underset{\mu_1 \geq 0}{\text{maximize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) \tag{C.12}$$

where

$$\Pi(S, \mathbf{Q}_b, \mu_1) = \sum_{d_i \in C_{1,h}}^{C_{\mathcal{B},h}} \sum_{k=1}^{K} q^k_{d_i, C_{j,h}} + \mu_1 \sum_j \sum_i \sum_k \left( s^k_{d_i, C_{j,h}} - (s^k_{d_i, C_{j,h}})^2 \right) \tag{C.13}$$

and the dual problem in (5.7) is given by

$$\underset{\mu_1 \geq 0}{\text{maximize}} \;\; \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) \tag{C.14}$$

The weak duality states that [16]

$$\underset{\mu_1 \geq 0}{\text{maximize}} \;\; \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) \leq \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \underset{\mu_1 \geq 0}{\text{maximize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) \tag{C.15}$$

For notational simplicity, let $\Omega(\mu_1) = \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1)$. Before going further with the proof, few things to note: i) $\forall \mathrm{S}, \mathbf{Q}_b \in \mathcal{W}, C4-1.3.1$, we have $\sum_j \sum_i \sum_k \left( s^k_{d_i, C_{j,h}} - (s^k_{d_i, C_{j,h}})^2 \right) \geq 0$, where $\mathcal{W}$ is the set of constraints $C4-1.1$-$C4-1.4$ and $C4-1.3.1$ ; ii) $\Omega(\mu_1)$ is a monotonically increasing function with respect to $\mu_1$; and iii) $\Omega(\mu_1)$ is bounded from above by the optimal value of primal given in (C.12). Assume that the dual problem given in (C.14) has the optimal solutions denoted by $\mu_1^*$ and $\beta^* = \{\widetilde{\mathrm{S}}, \widetilde{\mathbf{Q}}_b\}$, where $0 \leq \mu_1^* \leq \infty$. We now study the primal and dual problem for two cases. Assume that problem **P4-1** given in (5.7) has the optimal value give as $\gamma^*$, and accordingly $\gamma^*$ is also the optimal value of the primal problem given in (C.12). Therefore,

**Case I:** $\sum_j \sum_i \sum_k \left( s^k_{d_i, C_{j,h}} - (s^k_{d_i, C_{j,h}})^2 \right) = 0$

Under this case, the optimal solution $\beta^*$ is in fact a feasible solution for problem **P4-1** given in (5.7). Substituting $\beta^*$ into problem **P4-1** in (5.7) yields

$$\Omega(\mu_1 *) = \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1^*) = \Pi(\widetilde{\mathrm{S}}, \widetilde{\mathbf{Q}}_b) \geq \gamma^* \tag{C.16}$$

Since $\Omega(\mu_1 *) = \underset{\mu_1 \geq 0}{\text{maximize}} \; \Omega(\mu_1)$; thus, using (C.14) and (C.16), we can conclude that both the primal problem in (C.12) and the dual problem in (C.14) are in fact equal when $\sum_j \sum_i \sum_k \left( s^k_{d_i, C_{j,h}} - (s^k_{d_i, C_{j,h}})^2 \right) = 0$. That is,

$$\underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \underset{\mu_1 \geq 0}{\text{maximize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) = \underset{\mu_1 \geq 0}{\text{maximize}} \;\; \underset{\mathbf{S},\,\mathbf{Q}_b}{\text{minimize}} \;\; \Pi(S, \mathbf{Q}_b, \mu_1) \tag{C.17}$$

Another conclusion that can be drawn from the monotonously of $\Omega(\mu_1)$ with respect to $\mu_1$, we have

$$\Omega(\mu_1) = \gamma^*, \qquad \forall \mu_1 \geq \mu_1^*, \tag{C.18}$$

and accordingly, the results of Proposition 1 follows.

**Case II:** $\sum_{j}\sum_{i}\sum_{k}\left(s^{k}_{d_i,C_{j,h}} - (s^{k}_{d_i,C_{j,h}})^2\right) > 0$

Notice that, under this scenario, the dual optimization problem in (C.14) becomes unbounded from above as the solution of the maximization part with respect to $\mu_1$ is $\infty$ due to the monotonicity of $\Omega(\mu_1)$. However, this contradicts inequality of the duality defined in (C.15). Accordingly, **Case II** is not feasible. Therefore, for the optimal solution of problem **P4-1** given (5.7), **Case I** holds and the results of Proposition 1 follows immediately. ■

## C.4 Proof of convergence of Algorithm 1

We show that Algorithm 1 can be used to solve problems **P7-1** and **P7-2** and that the algorithm converges to a KKT stationary solution. By proving that Algorithm 1 converges, it can be in turn inferred that the *PAUSE* algorithm also converges.

*Proof:* The proof of convergence of Algorithm 1 has two aspects: the convergence of the algorithm to a solution and the nature of the solution.

**Convergence:**

Due to the linear nature of terms $f_b(\mathbf{Q}_b, \mathbf{S})$ and $f_a(\mathbf{Q}_a, \mathbf{V})$, their minimization results in an iteratively monotonically decreasing sequence. That is, at the $t^{th}$ iteration, we have

$$f_b(\mathbf{Q}_b^{t+1}, \mathbf{S}^{t+1}) \leq f_b(\mathbf{Q}_b^{t}, \mathbf{S}^{t}) \tag{C.19}$$

$$f_a(\mathbf{Q}_a^{t+1}, \mathbf{V}^{t+1}) \leq f_a(\mathbf{Q}_a^{t}, \mathbf{V}^{t}) \tag{C.20}$$

which follows from the fact that we use gradient descent to find the next solution point that results in the minimization of the objective function. Furthermore, as the constraints are reformulated to make the problem convex, finding a solution that results in the minimization of the objective function is attainable using typical convex optimization methods [16]. Consequently, we can conclude that there exists a cluster point of sequences $f_b(\mathbf{Q}_b^{t}, \mathbf{S}^{t})_{t=0}^{\infty}$ and $f_a(\mathbf{Q}_a^{t}, \mathbf{V}^{t})_{t=0}^{\infty}$ that result in the convergence of Algorithm 1 for a sufficiently small $\epsilon$, where $\epsilon$ is the stopping criteria defined as $||\widetilde{\mathbf{Q}}_b||_1 - ||\mathbf{Q}_b^{t}||_1 \leq e$ and $||\widetilde{\mathbf{Q}}_a||_1 - ||\mathbf{Q}_a^{t}||_1 \leq e$ respectively, with $f_b$ and $f_a$ denoting the objective functions of problems **P7-1** and **P7-2** respectively.

**KKT solutions:**

Considering the above discussion on convergence, let $(\bar{\mathbf{Q}}_b, \bar{\mathbf{S}}) \triangleq \lim_{t\to\infty}(\mathbf{Q}_b^{t}, \mathbf{S}^{t})$ and $(\bar{\mathbf{Q}}_a, \bar{\mathbf{V}}) \triangleq \lim_{t\to\infty}(\mathbf{Q}_a^{t}, \mathbf{V}^{t})$ denote the optimal cluster point solutions of problems **P7-1** and **P7-2** respectively, returned by Algorithm 1 with a sufficiently small $\epsilon$. Then, we can show that both $(\bar{\mathbf{Q}}_b, \bar{\mathbf{S}})$

and $(\bar{\mathbf{Q}}_a, \bar{\mathbf{V}})$ are KKT stationary points of the original problems **P4-1** and **P4-2** respectively. Considering the properties implied by the cluster point nature of the solutions found using Algorithm 1, we have $(\mathbf{Q}_b^t, \mathbf{S}^t) = (\mathbf{Q}_b^{t+1}, \mathbf{S}^{t+1}) = (\bar{\mathbf{Q}}_b, \bar{\mathbf{S}})$ and $(\mathbf{Q}_a^t, \mathbf{V}^t) = (\mathbf{Q}_a^{t+1}, \mathbf{V}^{t+1}) = (\bar{\mathbf{Q}}_a, \bar{\mathbf{V}})$ with $t \to \infty$ when optimizing problems **P7-1** and **P7-2** respectively. Accordingly, given $(\mathbf{Q}_b^t, \mathbf{S}^t) = (\bar{\mathbf{Q}}_b, \bar{\mathbf{S}})$ and $(\mathbf{Q}_a^t, \mathbf{V}^t) = (\bar{\mathbf{Q}}_a, \bar{\mathbf{V}})$ , for $(\mathbf{Q}_b^t, \mathbf{S}^t)$ and $(\mathbf{Q}_a^t, \mathbf{V}^t)$ to be KKT stationary solutions, $(\mathbf{Q}_b^{t+1}, \mathbf{S}^{t+1}) = (\bar{\mathbf{Q}}_b, \bar{\mathbf{S}})$ and = $(\mathbf{Q}_a^{t+1}, \mathbf{V}^{t+1}) = (\bar{\mathbf{Q}}_a, \bar{\mathbf{V}})$ should satisfy the following KKT stationary conditions

$$\nabla f_b(\bar{\mathbf{Q}}_b, \bar{\mathbf{S}}) + \lambda_b(\omega_b(\eta_b(\bar{\mathbf{Q}_b}) - \zeta_b(\bar{\mathbf{Q}_b})) = 0 \tag{C.21}$$

$$\nabla f_a(\bar{\mathbf{Q}}_a, \bar{\mathbf{V}}) + \lambda_a(\omega_a(\eta_a(\bar{\mathbf{Q}_a}) - \zeta_a(\bar{\mathbf{Q}_a})) = 0 \tag{C.22}$$

where $\lambda_b$ and $\lambda_a$ are the Lagrangian multipliers for conditions $C7-1.1$ and $C7-2.1$ respectively. Notice that, all other inequality constraints of both problems are constant terms and hence their derivatives are zeros and thus they are not part of the KKT conditions. Also, as the Lagrangian terms of the objective functions of problems **P7-1** and **P7-2** yield constant values, their derivatives of are zero. Accordingly, a quick comparison between the KKT conditions of problems **P7-1** and **P7-2** and problems **P4-1** and **P4-2** leads to that they are the same at the cluster point solutions. Consequently, $(\bar{\mathbf{Q}}_b, \bar{\mathbf{S}})$ and $(\bar{\mathbf{Q}}_a, \bar{\mathbf{V}})$ with the associated Lagrangian multipliers, $\lambda_b$ and $\lambda_a$ are KKT stationary solutions to the original problems **P4-1** and **P4-2**.