OFFLOADING AND CONTENT CACHING IN 5G HETEROGENEOUS NETWORKS: A GAME-THEORETIC PERSPECTIVE

by

Nirzhar Saha

Dissertation submitted in fulfilment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

Department of Engineering
Faculty of Science and Engineering
Macquarie University
Sydney, Australia

July 2018

STATEMENT OF CANDIDATE

I certify that the work in this thesis entitled "Offloading and Caching in 5G Heterogeneous Networks: A Game-theoretic Perspective" has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me.

In addition, I certify that all information sources and literature used are indicated in the thesis.

. . . . . .

Nirzhar Saha

Dedicated to my family

ACKNOWLEDGMENTS

If thesis is a journey, then I should thank those who helped me to complete this journey.

First and foremost, I sincerely express my gratitude to my supervisor, Associate Professor Rein Vesilo, for his guidance throughout the thesis. Both his technical expertise and his willingness to share ideas enriched me a lot and helped to reach a high level of understanding. I would like to thank my friend Md. Shareef Ifthekhar, who gave me feedbacks countless times, which helped thought the course of thesis. I would like to thank my friend and colleague Audri Biswas, PhD for his insightful comments about my research. I would like to thank Dr Keith Imrie for helping with the formatting and proofreading of this thesis. I am very grateful to Macquarie University for the financial support in the form of scholarships, top-ups and work opportunities.

Last, but not the least, I am extremely fortunate to have experienced the love and support showered on me by my parents. I hope, I have made them proud. My brother, with whom I discussed about few technical details during my thesis, who have always provided encouragement to me. Finally, I thank my loving wife for her support, throughout this journey.

ABSTRACT

The continuous evolution of wireless networks leads to the next generation of wireless network development, dubbed the fifth-generation (5G) mobile wireless network. The forthcoming 5G mobile networks will support diverse mobile data traffic with a one millisecond transmission delay along with reduced energy consumption. In fact every aspect of ongoing 5G research is pointing towards a much better quality-of-service (QoS) provision than its predecessor technology, the currently operating fourth-generation (4G) wireless network. It has been widely claimed that 5G will finally provide the technological edge and the necessary infrastructure to support the so-called internet-of-things (IoT). Thus efficient spectrum usage or spectrum management for 5G is of paramount importance to 5G fruition. While venturing the possibility of accommodating new spectrum seems audacious at this point, it will be the spectrum sharing technologies such as heterogeneous networks (HetNets) at the heart of initial 5G deployment. HetNets are becoming increasingly common and rely heavily on the spectrum reuse concept to provide services such as offloading and content-caching. Indeed recent literature puts much emphasis on offloading and content-caching as promising 5G technologies and they will play a vital role in futuristic distributed resource management. Firstly, this thesis focuses on developing a novel offloading approach for 5G dense HetNets. Secondly, this thesis attempts to devise a novel small-cell oriented content-caching approach.

The main contribution of this thesis can be summarised in two parts. In the first part, a novel pricing strategy is derived which enforces the low data rate macro users to associate with small-cells. The objective was to increase the spectrum utilisation of small-cells, while preventing cross-tier interference to macro users who are located close to small-cells. Furthermore the proposed pricing algorithm acknowledges one of the shortcomings of the traditional received-signal-strength (RSS) based user-association where a user always selects the strongest base station, i.e. the macrocell and thus it becomes overloaded in the process. The pricing algorithm achieves our design objectives as stated above. An evolutionary game-theoretic analysis is provided to show, how the proposed pricing strategy can influence users to select small-cells. Another important aspect of our proposed pricing technique is that a macrocell can control its population share (PS) precisely. By tuning the rate-threshold in the pricing algorithm, a macrocell can have an optimum population share. The evolutionary game captures the essence of the proposed pricing scheme. The mathematical background is provided to show that a unique solution exists. In addition, it is proved that the price-based network selection strategy is evolutionary stable.

In the second part, this thesis turns its focus on developing a novel caching strategy that utilises small-cells. Existing caching strategies mainly rely on small-cells for caching smaller contents. For retrieving larger contents, existing algorithms either download them from cloud storage or redirect an individual's request to a macrocell. Both approaches result in a higher delay, which motivates us to propose a caching strategy that enures more efficient utilisation of small-cells. The first caching model considers the cache-association strategies of users, which is modelled by a binary decision variable. Unlike existing work, this thesis formulates the small-cell based caching problem into a Nash bargaining game (NBG)

and utilises the cooperative nature of an NBG to devise a unique, fair and optimal caching strategy. However, the optimisation problem formulation which retains the Nash's axiomatic conditions, turns out to be an integer programming problem. Therefore constraint relaxations are applied to find a centralised solution, which can be used as a lower bound to derive the optimal solution. Afterwards Lagrangian relaxation to used to provide a distributed caching solution. In addition, a low complexity heuristic solution is proposed as an alternative solution to the centralised and distributed solution. By applying the concepts of two-dimensional coordinate geometry it is proved that the bandwidth allocation mechanism of the heuristic algorithm have has the same convergence property as Newton's method. Further investigation suggests that the bandwidth allocation mechanism approaches convergence in an iterative manner, therefore it exhibits a much desired property of distributed implementation and parallel computation. This thesis concludes by extending the previously proposed caching model including a cache-placement decision and providing appropriate solution methodology. The new NBG model involves a bilinear term and thus becomes a non-convex integer programming problem. The first step towards solving the problem is to apply a relaxation technique which will convexify the non-convex NBG by relaxing the bilinear term. McCormick relaxation is a well-known method used for convexifying an optimisation model involving bilinear term, which is used in conjunction with Lagrange relaxation to find a feasible solution.

Contents

List of Figures

List of Tables

List of Publications

Publications where the author is appeared as first author.

- N. Saha, R. Vesilo, "An Evolutionary Game Theory Approach for Joint Offloading and Interference Management in a Two-Tier HetNet", *IEEE Access*, vol. 6, pp. 1807 - 1821, December 2017.

- N. Saha, R. Vesilo, "Bargaining for Cache: A Game-theoretic Approach for Cooperative Content Associaiton in Heterogeneous Networks", to be submitted in *IEEE Access*, pp. 1 - 15, July 2018.

Chapter 1

Introduction

The idea of wireless communication predates from the early nineteenth century with a series of experiments in wireless telegraphy. It was James Clerk Maxwell who first theorised that electromagnetic waves could propagate through free space, which was later verified by Heinrich Hertz through a set of experiments in 1880 and 1887, thus laying the groundwork for Marconi's development of wireless telegraphy. However, Alexander Graham Bell is believed to be the first scientist to transmit data over a wireless link successfully by modulating light waves in 1880 [1], using a device called the photophone. In spite of the early inception of wireless signal transmission via the electromagnetic spectrum, it took more than a decade to deploy the first wireless communication system commercially. However, the modern digital telecommunication system was deployed in the late 1990s and evolved rapidly in a short span of time, delivering phenomenal technology advancements that have changed the way humans communicate and interact.

The meteoric rise in the popularity of media-rich devices (smartphones, tablets etc.) has led to a plethora of multimedia content-based innovations. Modern-day mobile users require 24/7 connectivity from their mobile network operators (MNOs), providing both connection-centric communication and content-centric communication. Both the first-

generation (1G) and the second-generation (2G) communication systems were primarily designed for providing connection-centric communication, i.e. end-to-end voice transmission. Mobile applications such as Skype, WhatsApp, YouTube, netflix etc. are examples of content-centric communication, causing a paradigm shift from connection-centric communication. These applications involve huge volumes of mobile data and are the main reason for the world-wide soaring traffic demand. As a consequence it is challenging the contemporary macrocell-based centralised wireless network architecture, in which macrocells are the focal point of all kinds of wireless services. The next generation of wireless network must have a decentralised architecture where small-cells will form a heterogeneous architecture along with macrocells and will play a crucial role in easing the macrocell burden.

A Cisco survey [2] sheds some light on existing and future trends of mobile data usage. It forecasts that wireless data traffic will increase up to sevenfold between 2016 and 2021. It also reports that global mobile data traffic grew about 63% in the year of 2016 compared to 2015, reaching about 7.2 exabytes per month at the end of 2016. One of the most intriguing statistics provided by Cisco is the rise of new mobile devices and mobile network subscriptions in 2016. Smart devices and internet-enabled machine modules have jointly accounted for half a billion new devices and connections in 2016. Thus a growing trend can be noticed where mobile-technology enabled devices are becoming increasingly common [3] and causing a paradigm shift towards the interconnectedness of all things, also called the internet of things (IoT). IoT-based emerging services have taken the communication concept beyond personal communications. The challenge therefore for the next generation of wireless networks is to manage a large volume of connected devices as well as the data flow stemming from them. Intuitively, managing this large number of devices is not only complex but also infeasible. Anticipating that the content-based mobile traffic demand will be increasing in the future, efficient offloading and handover

mechanisms will be of paramount importance to divert macrocell traffic (both content and cellular traffic) to small-cell networks. Novel insights are required to devise content-caching techniques so that the small-cells can host a large number of popular contents at the network edge and deliver them seamlessly upon receiving requests.

Wireless communication technology has evolved over the years to keep pace with the daunting demand of modern mobile networks. For example: the fourth-generation (4G) mobile communication standard is an upgrade from its predecessor third-generation (3G) mobile communication standard; it supports high mobility and can provide up to 100 Mbps with high mobility, and 1 Gbps in a low-mobility scenario, whereas 3G can provide up to 5 Mbps with a low-mobility scenario and 348 kbps in a high-mobility scenario. However, the theoretical bound for the data rate, determined through industry and academic research, has not yet been reached in a practical scenario. For example, the goal of 100 Mbps peak data rate per user has not yet been achieved in a commercial scenario [4]. Considering the trend of data explosion in recent years, the general consensus among the research community is that the area capacity or aggregate data rate will need to increase approximately 1000-fold from 4G to the next generation of wireless networks. To provide a seamless user experience the edge data rate and the peak data rate of a user need to increase to 100 Mbps and tens of Gbps respectively [5]. Massive-MIMO and mm-wave are considered as candidate technologies for data rate improvement. However both technologies are still in the development phase, and it will be the spectrum sharing technology of such dense heterogeneous networks (HetNets) which will provide the expected data rate improvement in the next generation of wireless networks.

Along with the data rate, the latency and energy-efficiency have become very important in the context of providing on-the-go user connectivity. A state-of-the-art 4G system embodying long term evolution (LTE) technology can deliver data with approximately 25 milliseconds delay, which exceeds the 1 millisecond delay requirement needed to en-

able real-time context-based applications such as wireless gaming [6], [7]. The round-trip delay requirement for the future wireless network is very stringent. The genre of delay-constrained futuristic applications is termed *tactile internet* which involves real-time human interaction in the fields of education and sports, entertainment, gaming, robotics and manufacturing, traffic, health care, the smart grid etc., and requires a round-trip latency of 1 millisecond for tactile steering and control of real and virtual objects without creating cybersickness. In order to fulfil the stringent delay requirement, small-cells must play a crucial role to share the burden of the macrocell load since small-cells generally have unused resources. Efficiently allocating these resources among a proportion of macrocell users can overcome the crucial problem of macro-network congestion. SBSs have very little congestion and locally retrieving contents reduces transmission delay. An efficient offloading mechanism will be able to divert a significant portion of users to small-cells with a view to reducing congestion and cross-tier interference. Considering the massive growth in mobile data consumption and the unprecedented increase in the number of networked mobile devices, the future wireless communication technology requires to be sustainable and withstand the growth and provide much-needed stability. To accommodate the demand stemming from human-centric communication and machine-centric communication, the fifth-generation (5G) mobile communication system has been envisaged as the future wireless network [8], [9].

1.1 5G Concept and Its Requirements

1.1.1 5G concept

As 4G has already been deployed and reached its peak in terms of the technological advancement it could provide, a fundamental question arises: *what is the next-generation wireless network going to be?* Badoi et al. have outlined the prerequisites of 5G concept

in [10], and cited a very high data rate as a primary requirement. They have proposed 5G technology as a combination of 4G along with functionalities which can bring unlimited wireless-world interconnection, convergence and cooperation, and capable of providing a variety of multimedia services at extremely high data rates. The aim is to design and develop next-generation wireless infrastructures that provide: i) a very high data rate, ii) a platform to globally integrate and inter-connect into a dynamically operating architecture. The main differences between 4G and 5G are outlined in [10] and [11]. The 5G terminal will be equipped with intelligence capability. It will be able to select the best-possible wireless access technology via information collection and processing about its surroundings. The 5G terminal will be able to combine different incoming streams from different technologies as well as support sophisticated error-correction and modulation schemes. The application layer in the 5G network must support an intelligent quality of service (QoS) requirement. Here intelligent QoS refers to the ability of base stations to adapt and adjust to changes in the radio environment so that QoS requirements in future wireless networks will always be satisfied. The 5G network must also support high-security protocols such as an authentication, authorisation and accounting (AAA) protocol, integrated between multiple systems [5]. In summary, the upcoming 5G network aims to meet the following challenges [12]:

- *Wider coverage*: To provide a ubiquitous high-speed service experience for users 24/7.
- *Massive capacity*: To meet the extremely high data rate requirement for each user.
- *Peak spectral efficiency*: 30 bps/Hz in the downlink and 15 bps/Hz in the uplink.
- *Massive connectivity*: To provide up to one hundred million connections.
- *Low latency*: To provide millisecond end-to-end delays.

- *Energy efficiency*: To support a high sleep ratio and long sleep duration to enable low energy consumption.
- *Dual connectivity*: To support macrocell and small-cell connectivity simultaneously.
- *Tactile internet*: Real-time interactive internet services.

1.2 Key Technologies in 5G

1.2.1 Spectrum Sharing Technologies

5G technology will face a very high data rate demand stemming from strong business requirements and much improved user experience compared to 4G networks. Considering that the radio frequency (RF) spectrum is limited, the 5G spectrum range will most likely be expanded towards the mmWave and up to the visible light frequency range. [12] states that 5G will be the starting point of the *full spectrum era*, where the possible frequency range will stretch from 1 GHz to 100 GHz. On the other hand, the authors in [13] explain how visible light communication (VLC) enabled small-cells, also called attocells, can be incorporated into 5G network architectures. The network capacity enhancement via VLC as part of a HetNet is analysed in [14], [15], and [16]. These research activities signify the importance of an extended spectrum for future wireless networks. However both mmWave and VLC are in their early stages of development and will definitely feature at some point in 5G deployment or beyond 5G development. Consequently spectrum sharing and spectrum reuse have appeared to be the primary candidates for providing data rate enhancement in 5G networks and will play a pivotal role to delineate 5G data rate and latency promises.

Spectrum sharing techniques have been developed to ensure efficient use of the precious radio spectrum. The advantage of spectrum sharing techniques is that they can benefit

from the existing infrastructures and network architectures, combining with intelligent resource allocation strategies to provide much needed data rate enhancement compared to new wireless access technologies such as VLC and mmWave. To enable spectrum sharing in 5G, the following technologies will be embodied in 5G networks.

Ultra-dense HetNets: State-of-the-art wireless technologies are already operating approximately around the Shannon limit [17]. Therefore limited capacity gain can be extracted with existing cell structures and frequency allocation techniques. As a counter-active measure, low power base stations (BSs) have been proposed to be deployed inside conventional macrocellular networks [18], exploiting both spectrum sharing and spectral reuse to provide localised services to users. It has been proved [19] that the densification of base stations by increasing the number of base stations per square kilometer, can increase the spectral efficiency (SE) of served users with a certain user outage constraint. Indeed, cell densification has been earmarked for generating additional capacity gain for a 5G wireless network [20] by utilising a three-dimensional capacity model (bits per second $\times$ Hertz $\times$ cells per square kilometer) [21]. Small-cells, e.g. microcells, picocells, femtocells and relays are categorised as low power BSs and are generally deployed inside a macrocell to form the HetNet (please refer to Fig. 3.1 for HetNet illustration). In a dense HetNet, a large number of small-cells are deployed in per square kilometer [17]. Note that small-cells have been deployed sporadically so far in 4G networks. For example: major mobile operators worldwide (AT&T, Verizon Wireless in USA; Korea Telecom, SK Telecom in South Korea; T-mobile, Vodafone in Europe; NTT DoCoMo in Japan; and Unicom in China) have started to roll out small-cell services to their clients. However they are being expected to deploy in an ultra-dense manner when 5G networks are finally deployed. Thus small-cells provide *ready-made* infrastructure for network expansion. Most importantly, underlay spectrum access based small-cell deployment enables spectrum sharing by taking advantage of aggressive spectrum reuse and embodying sophisticated interference avoid-

ance mechanisms [22]. Due to the availability of redundant resources in small-cells [23], it is possible to increase the per user data rate by redirecting a proportion of users from macrocell to small-cells. Such a data rate improvement is possible since users in each tier can be allocated higher bandwidth. In spite of the potential of dense HetNets, three factors will decide the success in harnessing capacity gain in 5G HetNets. They are: i) an efficient offloading and content caching mechanism, ii) interference coordination and management, iii) the access mode of its deployment. The primary functionality of small-cells is to provide localised communication. Therefore in 5G networks, small-cells are going to share the burden of macrocells by embodying efficient load balancing techniques such as offloading and content caching [20].

Offloading: The term offloading generally refers to the provision of alternative mobile data traffic connections to a wireless network such as either WiFi or small-cell, which is supposed to transmit through the cellular network [24]. Offloading has been earmarked as a key enabling technology to increase spectrum utilisation in 5G [20]. It can provide an inexpensive solution to the so-called spectrum scarcity for MNOs. However, the success of harnessing additional SE gain by deploying an ultra-dense HetNet largely depends on an efficient offloading algorithm. The mobile traffic in small-cells is more volatile, i.e. it changes rapidly than in macrocells. Moreover small-cells generally have unused resources. As small-cells are deployed by end-users, they provide *ready-made* infrastructure for MNOs for offloading macrocell users to small-cells. The most commonly used offloading strategy relies on the received signal strength (RSS) of the nearest base station. However RSS-based offloading can overload a base station, since a user always selects a base station which transmits with a higher power level (macrocells). Thus an efficient offloading algorithm is very important to harness additional SE gain from the small-cells. Most importantly, offloading will allow MNOs to deal with the increasing number of devices by distributing users in the lower tier.

Content Caching: Advances in the mobile internet continue to revolutionise our everyday life and are about to put their footsteps in the newest frontiers of human society such as health-care, electrical grid systems, robotics, industrial automation etc. Along with the new ones, the majority of modern-day mobile internet services are based on content. The emerging 5G networks will support these services, which also require real-time human machine interaction. In [25], this upgraded version of the mobile internet is defined as the tactile internet which will provide a further paradigm shift from content-centric communication. However the tactile internet requires a robust communication infrastructure with characteristics such as low latency, short transit time, high availability, and reliability. Most importantly it requires computing capability[1] to be implemented at the network edge [25], such that the end-to-end delay remains limited to one millisecond. It is expected that the 5G network will support these requirements, since current 4G systems do not support such strict requirements. Recently a lot of emphasis has been given to content caching via the mobile edge. Inspired by the resource availability and local storage capacity of the mobile radio nodes at the network edge, content caching via small-cells is considered as a lucrative option in terms of network performance improvement and economic incentives for operators. Both cloud cache and local cache have been proposed to deliver contents near to a user. The benefit is three-fold: i) this reduces the content retrieval time significantly, ii) it is energy efficient, iii) load balancing is achieved throughout the HetNet. Small-cell base stations (SBSs) are being densely deployed and this provides readily available infrastructures for content caching to be implemented. Consequently, caching via the network edge is going to be a prime feature of 5G mobile communication systems. However, the most important challenge in SBS-based content caching is to devise an optimal caching policy that maximises content retrieval at the network edge.

[1]Here, the term computing capability refers to an ability to compute an optimal solution of a sophisticated algorithm.

1.2.2 Network Intelligence

With the surge in user applications and a myriad of internet-enabled devices, along with ultra-dense HetNets, the network components of 5G systems certainly dwarf the size of current 4G systems. The network complexity of 5G systems is too multifaceted for engineers to comprehend. Managing such large-scale wireless networks is certainly not conceivable for engineers. In addition, the CAPital EXpenditure (CAPEX)[2] and OPerating EXpenditure (OPEX)[3] increase proportionally with the network size, which imposes financial constraints for 5G networks to become commercially viable. Therefore to provide better QoS and capacity, and to manage the network efficiently, network intelligence is touted as a much needed feature for the 5G wireless network.

In recent years, the research on implementing intelligent algorithms into wireless networks has intensified. The necessity of intelligence and autonomous function enabled wireless base stations has arisen from the complexity of managing a large number of base stations and users. Here the term users does not only refer to human subscribers but also extends to internet-enabled devices, capable of communicating with base stations or other devices. This emerging network architecture makes the classic approach of manual optimisation inefficient and time consuming, since human intervention is required for troubleshooting and system recovery. Furthermore, the legacy resource allocation procedure, which is currently being used, lacks both flexibility and a mechanism to adapt dynamically to its surroundings by taking advantage of the spatio-temporal and frequency variation of the radio environment. As a consequence, the achievable capacity and QoS of the legacy design results in sub-optimal performance. In summary, the existing resource

[2]CAPEX is defined as an investment cost, which is usually considered during the initial network roll-out phase or when the network is upgraded. It includes costs related to radio base station, transmission equipments, antennas, cables, installation cost etc. [26].

[3]OPEX constitutes the accumulated running costs of a network. It usually consists of site rental cost, backhaul lease, operation and maintenance costs of a mobile network [26].

allocation and management strategy lacks two key attributes, namely scalability[4] and agility[5]. On the other hand intelligent algorithms generally exhibit those aforementioned attributes [27] and therefore should be embodied in the 5G network.

The evolution of network intelligence is broadly divided into adaptive networks, autonomous networks, cognitive radio networks (CRN) and self-organised networks (SON). A large array of work [28], [29], [30], [31] has been done on adaptive networks[6] and autonomous networks[7], which are different from the legacy solution. Both networks respond to a situation based on predefined thresholds such as RSS threshold, minimum data rate, delay threshold etc., and lack the necessary intelligent mechanism to respond to a situation which does not belong to those predefined conditions [27]. Furthermore the learning capability which includes the ability to learn about the surrounding radio environment, e.g. signal strength of base stations, data rate etc., is a necessary tool which both adaptive networks and autonomous networks do not possess. Reference [27] argues that adaptive networks and autonomous networks lack scalability and agility, and sometimes they do not exhibit stability. In summary, both adaptive networks and autonomous networks do not exhibit the characteristics of scalability, agility and stability simultaneously to be categorised as SO algorithms. However both adaptive and autonomous algorithms lay down the foundation for both cognitive radio (CR) concepts and self-organising (SO) algorithms, which can learn and intelligently adapt to changes in the surroundings and can provide scalable, stable and agile solutions. CR is a widely researched technology for

[4]Scalability simply means that the solution complexity should not increase unboundedly with the system size.

[5]Agility refers to the responsiveness of a solution to changes around its operational environment while maintaining system stability.

[6]Adaptive networks trigger changes to their configuration in order to respond to changes such as signal-strength variation, data rate variation, pricing etc. in the system, when certain fixed conditions are fulfilled [27].

[7]Autonomous networks are adaptive networks and they do not require any external intervention [27].

acquiring idle and under-utilised spectrum pockets in the RF spectrum domain. A CR enabled device can intelligently collect and analyse under-utilised spectrum via a process called spectrum sensing. Thus the technology makes it possible to reuse spectrum dynamically and hence increase the wireless network SE via efficient spectrum sharing. The overall procedure includes spectrum sensing, spectrum allocation, spectrum access, and spectrum handoff. On the other hand, SO algorithms are adaptive in nature, and capable of simultaneously learning the radio environment by observing nearby interactions and independently deciding on a suitable action. Therefore the intelligent algorithms for 5G wireless networks fall in the category of either SO and CR based algorithms and can simply be referred as artificial intelligence (AI).

The authors in [32] explain the requirement and necessary application for an intelligent network algorithm to be implemented for infrastructure sharing and QoS management in 5G systems. They outline that agile system behaviour is important to reduce cost and improve user experience. They also propose that a learning capability should be implemented along with context recognition[8]. Reference [33] analyses possible candidate algorithms such as optimisation, reinforcement learning, game theory etc. for 5G technology, which fall into the category of CR or SO based implementation. They also discuss the problems where these solutions can be applied such as in offloading, inter-networking of the HetNet etc. The authors in [34] propose that the algorithms used for achieving SO should provide proactive solutions to satisfy the 5G delay requirement. They also emphasise combining network-level and cellular-level data to devise proactive load balancing, intelligent caching and preemptive traffic steering.

[8]Context recognition is a process of identifying a user's real-time contextual situations by analysing real-time wireless channel information using signal processing and machine learning algorithms.

1.3 Thesis Motivation

1.3.1 Offloading

5G mobile networks are envisioned to support diverse mobile data traffic with reduced energy consumption and improved QoS provision. A large proportion of the capacity gain in 5G networks will come from the utilisation of the lower tier of a HetNet, i.e. SBSs. Research directions naturally move towards technologies that exploit spectrum sharing and spectrum reuse of SBSs. Offloading thus becomes a viable solution which achieves both objectives by taking advantage of unused spectrum resources. Regardless of other technologies, those are going to be adopted in the quest for supporting a high data rate for each user; an efficient user offloading mechanism is of paramount importance to determine BS association with a particular SBS or the macrocell base station (MBS) with a view to maximising the per user SE.

After surveying the existing literature, we have found that several issues which hinder offloading techniques from slotting straight into 5G systems. Firstly interference (cross-tier and co-tier) is a well-known phenomenon which reduces the overall SE of a HetNet. Though offloading users to SBSs reduces cross-tier interference, co-tier interference may arise when a large number of users are offloaded to nearby small-cells. As underlay spectrum access based deployment is preferred over dedicated frequency channel deployment, and a prospective offloading algorithm must address interference issues which may be suffered by users in the macro-tier and small-cell tier. SBSs therefore must not generally accept users who are far away from them to limit co-tier interference. Furthermore, in some cases SBSs are known for creating interference hotspots, which are detrimental enough to create dead zones[9] for nearby macro users (MUs) in their downlink and uplink

[9]A dead zone is an area inside a cell where no mobile phone reception can be obtained due to the presence of interference, shadowing or blocking.

as discussed in [6], [7]. However the existence of SBSs can be exploited in a positive manner when a macrocell can avoid dealing with a serious cross-tier interference problem by offloading MUs to their serious interferer. This is indeed a win-win scenario for both macrocells and small-cells. Therefore joint interference management and offloading is necessary for future wireless networks.

Congestion is another key issue in the macro-tier which has a negative effect on maintaining a high level of SE for each user. Typically MBSs serve a large number of users compared to SBSs. It has been proved in [35] that SBSs have unused resources, since the small-cell load varies temporally depending on its subscribers' activation preferences. Small-cells, therefore, must share the burden of macrocells by accepting arbitrarily nearby MUs. In fact, dense small-cell networks can provide additional spectrum opportunities for MUs, assuming that they are deployed in hybrid access. Offloading thus comes in to the fray as a possible solution for congestion as well, where a portion of MUs join small-cells instead of macrocells to maximise their data rate. Motivated by the potential of offloading to provide a methodology that eases network congestion and interference in HetNet, the research objective is stated in Section 1.4, while a detailed analysis is presented in Chapter 3.

1.3.2 Content-Caching

Content caching is another technique by which a macrocell load can be distributed over small-cells. Note that both offloading and content caching fall into the category of load-balancing mechanisms. While an offloading mechanism typically redirects users to the small-cell tier irrespective of their connection request (either cellular or multimedia content), content caching via the network edge only accommodates users for the delivery of specific content requests. A few aspects that affect content- caching performance via the network edge are as follows. Firstly, existing caching policies under-utilise SBSs while

caching large data files [36], [37]. The upcoming 5G network standard has strict latency guidelines. Therefore an optimal caching policy should be devised emphasising locally available caching opportunities. Secondly, content diversity is another aspect which must be analysed when developing a caching policy. Content diversity means a large and varied category of popular contents, which should be delivered by SBSs rather than redirecting the majority of content requests to MBSs. Note that a full-fledged SBS-based caching policy must be able to cache as many popular contents as possible. While most of the literature emphasises video content delivery, the quality of videos has not received much attention. For example: a popular youtube video may have multiple video files for its content. What will be the caching policy in this scenario? Due to the storage constraint of an SBS, content diversity, i.e. different genres of contents, and placing multiple versions (for example; different versions of an audio/video) of the same content, will reduce the probability of placing other contents. Finally, the metric of fairness has not been considered when devising an SBS-based caching policy. Like any resource allocation mechanism, fairness is an important characteristic [38]. A large proportion of modern-day mobile data traffic originates from content sharing and content downloading. Fairness is therefore a key issue and should be introduced in a caching system to obtain an efficient solution. Motivated by these challenges the research objective is outlined in Section 1.4, while solution methodologies are presented in Chapter 4, Chapter 5 and Chapter 6.

1.4 Thesis Objective

Offloading

As has been mentioned before, the number of nodes (base stations or users) is going to be too large for a macrocell to manage 5G wireless networks. It is necessary to distribute a proportion of the macrocell load in an efficient manner to the small-cell tier. Therefore, it

raises a crucial question: how can a macrocell efficiently control its population share? In addition, shortcomings of the existing offloading algorithms further motivates us to design a novel offloading algorithm. The widely used RSS-based offloading scheme can overload macrocells [39]. In bias-based offloading schemes, users associated with small-cells can be subjected to unfavourable channels [40]. Note that biasing is a method of extending cell range by artificially increasing the signal strength at a receiver. Offloaded users can also suffer from interference from nearby macrocells [41]. In addition, both schemes lead to inefficient offloading, since they cannot guarantee that a certain number of MUs will always be offloaded to small-cells. Recently proposed game-theory-based network selection algorithms also cannot guarantee how many MUs will select a small-cell. The decisions of MUs on SBS selection largely depend on an SBS's price per unit of resource. Finally, macrocell pricing strategies lack control mechanisms, which induces offloading. To reduce macrocell burden and to achieve load balancing, a macrocell must have an optimal population share. Therefore an efficient algorithm for a macrocell is necessary which can offload users accurately to the small-cell tier. This motivates us to design a network selection algorithm based on a novel pricing function, enabling a macrocell to control its population share. The proposed algorithm guarantees that a certain percentage of the macrocell population, i.e. MUs, will be offloaded to small-cells as soon as the data rate per macrocell-associated user falls below a threshold. Furthermore, the pricing strategy of SBSs restricts distant MUs to select small-cells to reduce interference.

Content caching

After surveying the existing literature, we reached the conclusion that SBSs are vital elements of present and future wireless networks and therefore should be utilised efficiently to reduce the content-caching load on MBSs. For example: some contents can be exclusively cached at SBSs, or MBSs can apply a high price for specific contents to enforce

users to connect them from SBSs. Therefore, devising an SBS-based caching policy requires smart insights so that a significant proportion of users can be served locally for content caching service. The upcoming 5G network standard has strict latency guidelines, which strongly motivates us to design an optimal caching policy emphasising the locally available caching opportunities. Fairness is another issue, which motivates us to think differently when devising a caching policy. Fairness is an important feature when a large number of users compete for sharing a resource or when nodes allocate resources. For example: in a content-caching scenario, a large number of requests are handled by each SBS. Therefore an SBS should place contents in such a way that it provides a fair chance for users to access an SBS for content-retrieval purposes. The implication of fairness is that a fair cache-placement strategy can increase the cache hit rate at the small-cell tier. Users who are rational entities can compare individual allocations with others from their own point of view. Thus fairness may lead to a collaboration, which benefits both parties. In a competitive environment, fairness amongst users can reduce resource starvation and wastage [38]. This observation is extremely important for SBS-based caching systems, as users may reduce their delay requirement if the allocation of resource, i.e. bandwidth, is fair. Thus both the concept of fairness and the prospect of collaboration inspire us to propose a fair and optimal content-caching strategy via SBSs, and hence we can borrow Nash's axiomatic concepts and devise a content-caching problem in a 5G HetNet in the form of a Nash bargaining game (NBG) [42]. Our caching policy intends to answer the following questions:

1. How to increase SBS utilisation for a localised caching service?

2. How to incorporate fairness into the caching policy?

3. To develop techniques for avoiding a congested SBS for retrieving a content.

1.5 Thesis Contributions

In this thesis, we apply the pricing theory borrowed from economics [43] and game-theoretic approaches to enable offloading and cooperative content delivery in a 5G HetNet. The key contributions are listed as follows:

- **Pricing and evolutionary dynamics**: In Chapter 3, a novel price function for macrocells by combining two different pricing strategies, is formulated. Unlike existing pricing approaches, our proposed pricing strategy enables macrocell-controlled offloading in a dense HetNet environment. The small-cell density has been constructively exploited in a HetNet to harness offloading opportunities for MUs, the goal of which is to solve two problems simultaneously. It aims to shun the macro users, who have a lower received data rate due to the interference perceived from the small-cell tier. The scheme also fends off congestion in a macrocell, by offloading macro users to the small-cell tier. The price function has a *threshold pricing* strategy, which a macrocell adopts, with a view to influencing low data rate macro users to join a small-cell network. Small-cell networks also charge a price, which includes an access price and an interference compensation price, proportional to the number of macro users who choose the small-cells instead of the macrocellular network. To justify the efficiency of the pricing scheme an evolutionary game-based framework is formulated. This enables us to model and analyse the behavioural dynamics of the large number of MUs under the proposed pricing strategies of both networks. Replicator dynamics is used to find the evolutionary equilibrium of the evolutionary game. The proof of the existence and uniqueness of the evolutionary equilibrium is derived through extensive analysis. It has been shown that selection dynamics converges to the evolutionary equilibrium. By utilising the Lyapunov's direct method, it is also proved that the evolutionary equilibrium is asymptotically stable.

- **Bargaining game for cache association**: Chapter 4 presents a novel outlook on the content-caching problem for the 5G HetNet. One of the problems in SBS-based caching is to provide fairness to users who send requests for downloading a contents. To ensure optimality and fairness at the same time a Nash Bargaining model is applied in a cooperative caching scenario for a dense HetNet. The interaction among SBSs and users for content retrieval purposes is modelled in the form of a joint cache association and bandwidth allocation. Utilising the property of the NBG a unique solution is obtained which guarantees fairness and optimality in terms of ensuring the lowest average download delay per user.

- **Heuristic algorithm for cache association**: Chapter 5 presents a heuristic approach to solve the content-caching problem for the 5G HetNet. One of the problems in the NBG-model proposed in the Chapter 4 is that it comes with an inherent computational overload. As a result a heuristic algorithm is presented which solves the content-caching problem proposed in Chapter 4 in two stages, namely the association stage and the bandwidth-allocation stage. It is shown that the bandwidth allocation follows Newton's method for convergence and thus a faster convergence is guaranteed.

- **Joint cache-association and cache-placement optimisation**: Chapter 6 presents a content-caching scenario where the cache association depends on the content-placement decision of an SBS. The problem becomes more complex with the introduction of a second binary decision variable. Because of the presence of both binary variables in the constraint function and the objective function, the NBG-based optimisation problem turns out to be a bilinear non-convex integer programming problem. To preserve Nash's axioms, a convex relaxation method called McCormick relaxation is applied. Afterwards, the convexified problem is solved using

Lagrangian relaxation method. Finally, a subgradient update is used to find the optimal solution of the convexified problem, which provides a lower bound of the joint optimisation problem.

1.6 Thesis Outline

This thesis starts with a brief statement about the evolution of wireless networks to date, which gradually leads to the development of the 5G standard. Afterward, a synopsis of the key enabling 5G technologies such as HetNets, offloading and content caching is provided. The scope of the contributions in these technologies such as offloading and content caching works as a motivation for the proposed price based offloading mechanism in Chapter 3 and content-caching techniques in Chapter 4, Chapter 5 and Chapter 6. Chapter 2 discusses 5G standardisation activities and the research status of offloading and content caching briefly. The efficiency and the ease of resource management are the main motivating factors behind the proposed game-theory based offloading mechanisms and content-caching schemes. Therefore, in Chapter 2, Section 2.5 briefly discusses relevant game theories which are used as mathematical tools to formulate and analyse the proposed offloading mechanism and content-caching schemes. In Chapter 3, an evolutionary game-theoretic framework for a proposed price-based offloading mechanism has been presented. This chapter details the evolutionary network dynamics under the proposed threshold pricing schemes and provides an extensive analysis for the existence, uniqueness and stability of the evolutionary equilibrium. Chapter 4 proposes a novel cache-association algorithm with a view to increasing SBS utilisation for content caching. The chapter first provides a detailed analysis on acquiring a centralised solution. To satisfy the distributed architecture of 5G implementation, an intelligent scheme based on the subgradient method is also proposed. Chapter 5 presents a heuristic algorithm with a view to reducing the

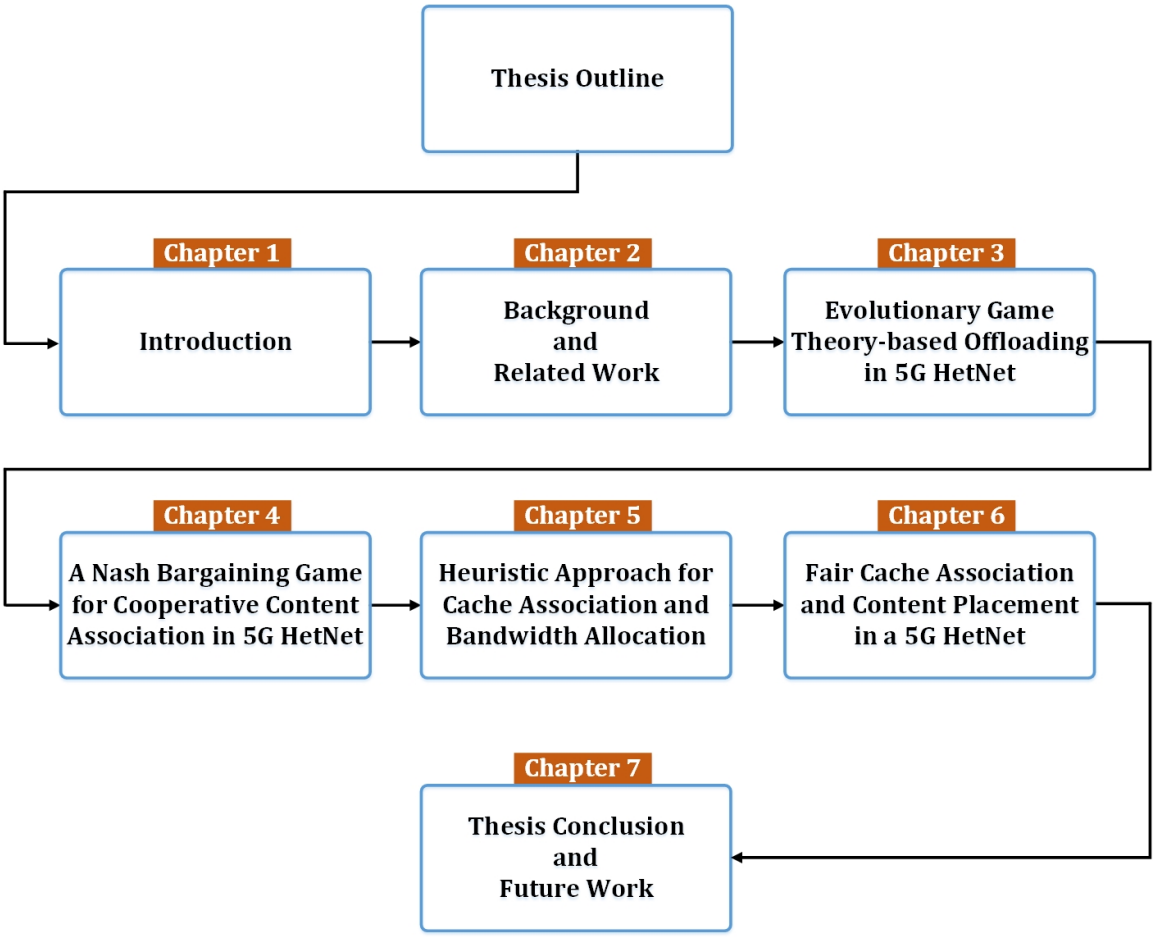


Figure 1.1: Graphical illustration of thesis outline.

computational overhead of the NBG-based optimisation proposed in Chapter 4. Chapter 6 proposes a joint optimisation of content association and content placement, formulated in terms of an NBG. Chapter 7 concludes the thesis by summarising its contributions regarding the offloading mechanism and content-caching schemes, as well as highlighting the scope of future work. The thesis outline is graphically illustrated in Fig. 1.1.

Chapter 2

Background and Related Work

2.1 Chapter Introduction

In this chapter, a brief introduction to 5G standardisation trends and research aspects is provided to coincide with the objectives of this thesis. Afterward, a comprehensive review is provided of notable research on two of the most promising 5G technologies, namely offloading and content caching. Finally, preliminary concepts of the well-studied game-theoretic algorithms such as an NBG, and an evolutionary game theory (EGT) are discussed along with their application. These game-theoretic concepts are fundamental to obtain analytical frameworks for the proposed problem formulations in Chapter 3 for price-based offloading, Chapter 4 for joint cache association and bandwidth allocation, Chapter 5 for decoupled cache association and bandwidth allocation, and in Chapter 6 for joint cache association and cache placement.

2.2 Standardisation Activities in 5G

Standardisation is the first step towards the commercialisation and deployment of 5G technology which will ensure a global standard with a key performance guarantee. There-

fore standardisation bodies are at work to provide a unified 5G standard. The ITU-R working party (WP) 5D is preparing a list of candidate technologies [44], which will ensure minimum performance criteria [45] during 5G deployment. Some notable minimum technical performance requirements are 20 Gbps download and 10 Gbps upload speed at the peak data rate, ultra-reliable and low-latency communications (URLLC) delay to be one millisecond for user-plane latency performance, high reliability in terms of success probability within the range of $1 - 10^{-5}$ for transmitting a layer-2 protocol data unit of 32 bytes within 1 ms delay at coverage edge, low energy consumption and at least 100 MHz bandwidth provision for efficient operation. The complete list of requirements can be found in [44]. The process of certifying a 5G standard of ITU-R is expected to be completed by late 2019. The European standardisation body, 3GPP, has also started a standardisation procedure by the publication of Release 14 in the year of 2017. A key outcome of Release 14 is providing an ability for a user equipment (UE) to support dual connectivity, i.e. simultaneous connections with a macrocell and a small-cell. Phase 1 of the 5G standardisation process 3GPP is going to conclude tentatively by September 2018 with the publication of Release 15. However, a complete 5G standard is expected to be finalised in the year of 2019 with Release 16, detailing 5G technology requirements along with its application for various services and scenarios, which aligns with the ITU-R standardisation time line.

2.3 Significance of HetNets in 5G

In the light of the requirements stated in Section 2.2, it can be concluded that the task of providing a unified platform meeting all these requirements is extremely challenging. Network densification, mm-wave and massive-MIMO are considered as instrumental to achieve the aforementioned goals and are termed the *Big 3* technologies [44]. The expected

capacity gain by means of mm-wave technology sounds promising. However the practical implementation of this technology is complicated for various reasons. Firstly, the free-space pathloss is very high due to the high range of frequencies of mm-wave. In addition, it is also a challenge to maintain link quality when a UE is moving at a high speed, because of the Doppler spread. Atmospheric attenuation and shadowing affect the link budget of mm-wave. Furthermore, the results obtained from outdoor measurements show that mm-wave links suffer from a higher outage probability. Indoor mm-wave links also suffer high outage probability because of shadowing and blocking [44]. Thus maintaining a stable performance and simultaneously meeting different types of QoS requirement become very difficult. massive-MIMO is another technology which is highly regarded for overcoming the capacity bottleneck. In massive-MIMO, performance improvement comes at a cost of complex signal processing. It requires complex pre-coding, very accurate positioning of the receiving antenna and sophisticated techniques for the mitigation of pilot contamination[1] [46], [47]. Theoretical bounds derived in a laboratory, which give an impression of a drastic performance improvement in terms of a high data rate and overall performance improvement, are yet to be achieved in a complex practical scenario such as in an urban environment. Therefore both the mm-wave and massive-MIMO technologies are lagging behind their other Big 3 counterpart, network densification.

Network densification, or cell densification, is now considered to be the initial driving force behind 5G deployment, while both mm-wave and massive-MIMO technologies will embody 5G base stations at some point after initial deployments. Small-cells are placed inside a macrocell coverage and thus the network architecture evolves into a HetNet. Small-cells were initially deployed by end-users as closed access for either improving sig-

[1]Pilot contamination is a phenomenon which arises in massive-MIMO technology due to the fact that a finite number of pilot sequences must be reused for the uplink channel estimation. As a consequence, each pilot sequence is corrupted by inter-cell-interference as well as noise when received at the base station [46], [47].

nal reception or to bridge the gap of the coverage holes inside a macrocell. The statistical data of the small-cell performance gain obtained from the initial deployment phase motivates MNOs to deploy small-cells in an ultra-dense manner in 5G HetNets. According to [48], more than 10^3 cells/km^2 constitute an ultra-dense HetNet and serve up to 600 active users/km^2. RF spectrum is scarce and expensive. Ultra-dense HetNets provide an expensive solution in the quest for higher SE. To take full advantage of dense HetNets, MNOs are now considering hybrid-access small-cell deployments [23] instead of conventional closed-access deployment. As a result, MNOs are looking to find a way to motivate privately deployed small-cells to relax their access strategy to hybrid mode [49] so that targeted macro users can be offloaded to their serious interferer. In addition, MNOs can deploy small-cells in public places to give to users having a high QoE. Thus we are moving towards a multi-tiered HetNet in the future, where the small-cells in the lower tier will provide various services leveraging an efficient offloading mechanism and content-caching schemes. The downside of such ultra-dense HetNets is an increased complexity in network management and higher inter-tier and intra-tier interference. However, with an efficient offloading mechanism, each small-cell in the HetNet can be viewed as an opportunity for offloading traffic, as it improves user QoS and the small-cell's spectrum utilisation significantly. Since SBSs are generally user-deployed, utilising hybrid-access-based SBS deployment can reduce CAPEX for MNO [26]. In addition, small-cells require less investment for installation [26], [50], [51]. Furthermore SBSs are equipped with self-organising capability and network automation features, which reduces the requirement of human intervention. As a result, MNOs benefit by the reduction of both CAPEX and OPEX, if small-cells can be utilised on a bigger scale [26], [51]. Thus small-cells are going to be extremely important in the context of a low-latency wireless network, by reducing end-to-end data-transmit time intervals. Therefore, with an efficient offloading mechanism and a content-caching technique, a paradigm shift is envisaged from a macrocell-centric

centralised cell structure to a small-cell-centric distributed cell structure in a 5G HetNet.

2.4 Related Work

2.4.1 Recent Advancements in Offloading

The role of small-cells in 5G networks has become more important than ever for managing interference and providing an inexpensive solution to increase SE utilising offloading. In the following, the recent research on offloading is summarised into two categories.

Game Theory Based Offloading

The authors of [52] categorise game-theoretic applications in wireless network selection based on the interaction among players into three groups: game between networks [53], game between users [54] and game between network and users [55]. Broadly the game-theoretic contribution for offloading-based network selection can be divided into non-cooperative and cooperative game theory. Non-cooperative game theory has been applied in [55], [56], [57], [58] for offloading users to base stations, i.e. redirecting users from macro-cells to WiFi and small-cells. In [56], the concept of cell-range expansion (CRE) is utilised to avoid cross-tier interference by offloading interference affected MUs to small-cells. A Stackelberg game is formulated to model the interaction of macrocells selling bandwidth to small-cells so that victim MUs can be served by the small-cells. A three-stage Stackelberg game framework is provided in [57], where spectrum leasing and hybrid accessing are jointly considered to offload MUs to femtocells. It has been shown that spectrum leasing and hybrid accessing jointly improve the utility of both tiers. The authors of [58] consider a non-cooperative game to analyse a user-offloading mechanism from an economic perspective. A multi-leader multi-follower data-offloading game is proposed to capture the interaction among APs and base stations. A similar approach has been adopted in [55],

where users select a network selfishly to minimise their access-selection cost. A two-stage multi-leader and multi-follower game is proposed to capture the interaction among APs and users. Both [58] and [55] compute the subgame perfect Nash equilibria to demonstrate the existence of the equilibrium in the respective cases. Note that the subgame perfect Nash equilibrium is a refinement of Nash's non-cooperative games and it represents a separate branch of modern economics in which a non-cooperative game is structured in multiple subgames. Details on the subgame perfect equilibria can be found in Chapter 5 of the book by Takako Fujiwara [59]. It should be noted that the Nash equilibria resulting from non-cooperative game-theory are always inefficient [60] and hence have a sub-optimal performance compared to cooperative games. The implication for wireless networks is that those non-cooperative offloading schemes will never be able to offload the optimum number of users. A cooperative capacity offload is proposed in [61], for cell-edge MUs, who suffer from throughput degradation due to cross-tier interference. These MUs are offloaded to SBSs to enhance their QoE. The joint base-station association and power-control problem is formulated as an NBG to perform capacity offloading. The cooperative NBG always leads to a unique Pareto-optimal and fair solution. The bargaining game is applied by the authors of [62] and [63], where the base stations act as players and they collaborate to provide a fair and optimal downlink association for users in HetNets. In [62], a data-rate based utility is adopted with a view to guaranteeing a certain data rate for users while maintaining fairness for all users. The authors of [63] consider an opportunistic user association algorithm, formulated as an NBG to maintain the QoS for multi-service traffic. They classify human-to-human traffic as the primary service and machine-to-machine traffic as the secondary service. The contribution is unique in the sense that the proposed opportunistic user-association scheme allocates resource in a fair way to the secondary service without affecting the primary service performance.

EGT-based network selection has been studied rigorously in [49], [54], [64] and [65]

with a view to reducing the macrocell load. An EGT-based network-selection algorithm is presented in [54] to avoid overloading a particular network, where users select any base station from an IEEE 802.16, or an IEEE 802.11, or a CDMA cellular network. A two-level game model is proposed in [49], in order to distribute the load (macro users) inside the coverage area of a HetNet. Each small-cell determines the amount of open-access bandwidth based on the incentive offered by the macrocell. Both networks then compete for macro users in the HetNet by declaring a price for their resources. This pricing interaction is then modelled using a dynamic Stackelberg game as an upper-level game. Users on the other hand choose either of the two networks based on the price offered by each network and the satisfaction level. EGT is used for modelling user behaviour in the HetNet as a lower level game to find the proportion of users who are associated with the different network tiers. Network selection is viewed as a user-association problem in [65]. A population game is considered for finding the proportion of users who will be distributed in a network of multiple base stations in order to maximise their individual payoff. Three different kinds of evolutionary dynamics are evaluated to find the dynamics which converges faster to an equilibrium. The authors of [65] develop a new algorithm called the best-responding user association with traffic estimation (BRUTE) by utilising the fast-converging property of the best-response dynamics of the population game.

Non-game Theory Based Offloading

User-association schemes akin to network-selection schemes have been proposed in [39], [41], [66] with a view to offloading MUs in HetNets. The conventional received signal strength (RSS) based user-association scheme [39] is not suitable for HetNets due to the variable power levels between macrocell base stations (MBSs) and SBSs. This scheme also does not consider the traffic load of a base station. Therefore, users will always connect to the base station with the higher transmit power in this process. As a consequence,

a large proportion of users will always end up associating with macrocells and overload them in the process. Furthermore, low-power SBSs, with their low antenna gain, will result in lightly loaded small-cells. In addition, there will always exist a mutual interference between macro users and small-cells. In order to overcome these inefficiencies of the RSS-based offloading technique, bias-based cell-range expansion (CRE) has been proposed in [66], [41], where SBSs artificially increase their transmission power by adding a positive bias to the transmitted signals. In [66], the authors have shown that the capacity of offloaded users to small-cells can be improved via biasing. Nevertheless users offloaded to biased base stations will suffer from strong interference from nearby macrocells [41], as well as unfavourable channels from biased base stations [40]. As a result, there exists an offset between the attainable performance achieved via offloading and the perceived interference from macrocells. In addition, the bias value needs to be optimised carefully, since there exists a trade-off between the number of offloaded users and the system throughput. Higher bias values may attract more MUs, which will in turn overload small-cells, and lower bias values may result in an inefficient offloading and severe degradation in achievable throughput per user. The authors of [67] propose a Q-learning approach to learn the bias value intelligently from past experience to minimise the service outage. However, [67] does not provide the insight about the effect of interference on users for bias-based cell association. The authors of [68] consider a distributed optimisation model which jointly utilises both cell biasing and almost blank sub-frame (ABS) based inter-cell interference coordination (ICIC) to efficiently offload users to small-cells and simultaneously mitigate interference to offloaded users. Firstly, the dual decomposition based solution method proposed in [68], does not converge fast enough and may cause performance degradation to users requiring real-time services. Secondly, offloaded users are still prone to interference from macrocells since macrocells are allowed to transmit during ABS. The authors of [69] considers both spectrum partitioning and offloading in a two-tier HetNet so that a certain

throughput is always guaranteed to offloaded users. A convex optimisation technique is adopted in [70] for dynamic user association to improve the SE of users with a view to maximising the sum data rate of users. However, it results in congested small-cells and lightly loaded macrocells. To deal with this issue, the authors of [71] propose a load-aware cell-association method for downlink HetNets, where MUs are pushed onto lightly loaded small-cell tiers even if those small-cells offer a small instantaneous signal to interference plus noise ratio (SINR). Note that a minimum SINR is always satisfied for users who associate with small-cells. A classic optimization theory is used to jointly formulate the cell-association and resource-allocation problem, where a logarithmic utility maximisation problem is solved by using a dual-decomposition technique. The logarithmic utility function imposes a proportional fairness criterion and thus users avoid resource starvation, unlike [70] where a significant proportion of macrocell users are starved of resource.

Summary

In conclusion, it is observed that the game-theoretic approach and the non-game-theoretic approach do not ensure how many users are going to be offloaded in the small-cell tier. Most importantly, macrocells lack an efficient *control mechanism* to guarantee that a certain proportion of the population will always be offloaded to the small-cell tier. Since the number of small-cells is going to increase, therefore 5G network will require an offloading mechanism which allows a macrocell not only to offload a fraction of its users to the small-cell tier but also to distribute them throughout the lower tier in such a way that small-cells also have an optimum population share.

2.4.2 Content Caching

Caching, by redirecting content-based mobile traffic towards SBSs, has emerged as a promising solution to avoid some of the traffic burden on MBSs [72], [73]. SBSs are low-

power base stations, generally deployed inside macrocells to form heterogeneous networks (HetNets) with a view to enhancing the user quality of experience (QoE). SBSs in HetNets are considered as an integral part of a fifth-generation (5G) wireless network to provide affordable options for enhancing the network capacity. They provide seamless localised communication to a group of users by utilising spectrum reuse. Inspired by its potential, user offloading via SBSs has been proposed as a possible solution to reduce congestion [74], [20] and can be utilised similarly for content caching. Deployed SBSs in macrocellular networks provide the necessary infrastructure for content caching, where the contents are stored by utilising the storage of SBSs. Since users are being served locally, the SBS-based caching technique reduces the per-user content-retrieval delay. In this way it brings a network topology shift on how contents are going to be delivered to end-users, unlike the traditional information-centric networking (ICN) where content providers (CPs) have to rely on third-party access points (APs) for delivering contents [75]. While we restrict our study to SBS-based caching, discussing the inefficiency of the traditional ICN-based caching remains out of the scope of this thesis. In an effort to summarise the main advantages of SBS-based caching, we list them as follows:

- Users submitting content download requests do not need to download data from a remote server since they are being served *locally*, which reduces the content-retrieval delay significantly.

- SBS-based caching negates the redundant transmission of popular contents over backhaul channels.

- It reduces both the infrastructure cost and the operational cost, while providing reliable service compared to wifi-APs.

- Finally, SBS-based caching allows the content traffic load to be distributed among multiple SBSs, prompting users to download content without facing significant con-

gestion.

Recent Advancements in Content Caching

Content-caching functionalities are preferred to be performed at small-cells in 5G networks. Two of the most fundamental questions associated with SBS-based caching are: *What factors should be considered for an optimal caching policy?* and *How can the content-retrieval latency be minimised?* The first question involves caching anticipated or popular contents for various groups of users. The challenge is to determine the popularity effectively. For example: contents can be placed in an SBS on the basis of their popularity to maximise the metric called the cache hit rate. From the technical point of view of a HetNet, a higher cache hit rate will reduce the traffic burden on macrocells. However the content popularity changes with time, which makes it difficult for an SBS to keep its content placement up-to-date since each SBS needs a global knowledge of the content pool along with the knowledge of content popularity. In an SBS-based caching scenario, users may request a content which is not popular in the global setting, i.e. a specific content is popular in a local setting but it is not popular in a global setting. The second question deals with both the placement issue of the cached content and the association rule for the users, e.g. storing contents in an SBS in a way that it will incur less delay to deliver upon receiving requests, or how to select a suitable SBS which is serving only a few users assuming that the SBS has the desired content. Therefore, in a 5G network the challenge of an SBS-based content-caching system is to deliver a large number of content from the lower tier satisfying the 5G latency requirements.

A centralised optimisation problem is devised in [76] to portray the cache placement problem. The authors also consider the popularity of any contents before placing them in an SBS. The metric 'popularity distribution' changes slowly and quickly for larger contents and smaller contents respectively. However, in this method a problem arises

when a new content suddenly gains popularity. Since it is assumed that the popularity of the larger contents changes slowly, users will have to download newly popular content from MBSs. In addition, from an economic perspective an MNO can in fact miss out on revenue either by not placing the content or by delaying the placement.

The total number of contents that can be cached at an SBS depends on its storage capacity. This imposes a new constraint on designing an optimal cache-placement problem. Wang et al. in [36] consider a two-stage caching policy to alleviate the storage constraint. Smaller contents are cached at the network edge, i.e. at SBSs, and larger contents are placed at the evolved packet core (EPC). Tran et al. in [37] propose a hierarchical caching concept that utilises a central cache manager to place contents and a global indexing table to keep track of geographical knowledge of the content placement. The central baseband unit acts as a cloud storage and contains larger contents. Smaller contents are stored and delivered from the network edge while keeping the total cache size of the network fixed. However in both cases users are affected by a significant delay since they have to fetch larger contents from the EPC [36] and from the cloud [37]. Both [73] and [72] propose single-stage caching that involves edge nodes only. The authors of [73] study the effect of wired backhaul as well as wireless channel quality for an SBS-based caching system. An optimisation problem is developed for joint caching and user association that aims to minimise the average download delay. They argue, and rightly, that a caching policy may fail in the absence of a suitable user-association rule so a user may experience delay due to congestion, especially when popular contents are cached at an SBS. In [72] an optimal cooperative content caching and delivery policy has been analyse in the form of an integer-linear programming problem (ILP) to design an efficient caching strategy to reduce the user-perceived delay. The authors investigate a hybrid caching strategy that utilises user equipments (UEs) along with SBSs as possible nodes for content caching. [77] presents a realistic observation where the authors investigate a distributed caching policy for layered

video content delivery in the form of an optimisation problem. This problem tries to model a practical scenario where a user may request a high-resolution video instead of the low-resolution video placed at an SBS. In this case, after receiving such a request, if the requested video with the preferred layer is unavailable at the SBS of interest it will retrieve the content from another SBS. Thus SBSs cooperate to overcome the storage constraint and devise a joint caching policy. Thus [77] briefly addresses the problem of content diversity by maximising content placement at SBSs via cooperation. However, a user perceives a higher delay since the requested content is fetched from another SBS by the SBS of interest, which adds to the transit delay. Thus an offloading mechanism is necessary in this case to avoid a larger delay.

Summary

It should be noted that the 5G wireless network has a stringent delay requirement [44], [45]. Therefore SBS-based content-caching technology is extremely important in the context of the 5G network to reduce end-to-end delay for the user and content-caching load on macrocells. However the challenge arises to host contents at the network edge which are locally popular, since the aggregate popularity of contents in a global setting may not match the content popularity in a local setting, which makes content placement for SBS-based caching difficult. Caching multiple contents at an SBS is difficult as well as impractical due to the cache-size constraint of each SBS. In addition, existing content-placement algorithms hardly consider a cooperative caching strategy which can ease the cache-size constraint of each SBS, where multiple SBSs form groups and judiciously devise a local content pool. Therefore SBS-based content caching requires an insightful approach with a view to reducing the macrocell caching burden .

2.5 A Brief Introduction to Game Theory

Game theory is a branch of applied mathematics which provides an analytical way of modelling the interaction of agents formally known as players. It can be said that a game is bring played between two individuals whenever they meet each other. From the football pitch to the table of negotiations, from social science to politics, from biology to modern wireless communication, whenever a strategic decision has been made, a game has been played. Game theory can mathematically map these interactions among several decision makers who can have conflicting or common interests, and provides a methodology to compute the outcomes of those interactions, either numerically or iteratively, which lead to a stationary solution, i.e. convergence. For this reason, classical game theory has been applied either to solve complex wireless networking and communication problems or to capture the dynamics of it [78], [79]. As modern-day wireless networks consist of a large number of wireless nodes (UEs, sensors, machines), we need a mathematical tool that allows implementation of SO functionalities. Game theory provides exactly this. Before discussing the application of game-theoretic approaches, a brief introduction to game theory is provided, where especially the NBG model and the EGT model are discussed.

The field of game theory can be broadly categorised into non-cooperative game theory and cooperative game theory. A non-cooperative game describes a competitive situation where self-interested players make strategic choices independently in a way that serves their own interest. The term non-cooperative does not always imply that the players do not cooperate, but it means that any cooperation that might arise must be self-enforcing, with no communication or coordination of strategic choices among the players [79]. For example: the decision making of intelligent wireless nodes such as MBSs and SBSs to control their transmit power is an example of a non-cooperative game. While increasing an individual transmit power has an incentive, the interference generated will have a negative effect on other users and may outweigh the benefits gained by the transmit power

increase. In contrast, cooperative game theory deals with the study of rational players' behaviour when they cooperate. For example: consider the aforementioned scenario where wireless nodes are trying selfishly to increase their power even though these nodes have a common conflict of interest. Here increasing the power is a common interest of wireless nodes. At the same time it is also a conflict of interest because increasing a node's power will affect another node in terms of interference. Instead of increasing power selfishly, they can come to an agreement to keep the power increments at a certain level so that interference remains at a tolerable level. Thus cooperative game theory resolves such conflict via cooperation. Both cooperative and non-cooperative games have been applied in wireless communication networks to solve resource-allocation problems. However, both cooperative and non-cooperative games are generally applied in a one-shot manner and therefore do not capture the dynamic nature of wireless networks, since both types of game do not permit updating strategies. As a result dynamic games such as EGT and differential games are currently being considered, especially to capture the behavioural dynamics of the users in a wireless network. In the following we keep our focus on the NBG and EGT, which we use rigorously in the following chapters.

2.5.1 Cooperative Games

Bargaining-based game theory is a special branch of cooperative game theory which provides analytical pathways to study rational players' behaviour when they cooperate. It describes the bargaining phenomenon of players in a conflict and how they reach an agreement that is in their mutual interest [42], such as bargaining over the division of a certain resource. However each player remains interested in holding out for the most favourable settlement possible. In the following subsection, we introduce Nash's axiomatic approach for reaching a bargaining solution.

Nash Bargaining Solution

An axiomatic approach for obtaining a solution of a bargaining game was introduced by John Nash in 1950 [80]. It provides an analytic framework to capture the interaction among multiple entities, referred to as players, who are trying to reach an agreement over trading or sharing of certain resources but have conflicting interests about the method and terms of reaching that agreement. Thus the solution of a bargaining game mutually benefits the players who remain involved in the bargaining process but have no conflict of interest on the terms of agreement. A bargaining game represents a set of outcomes expressed in terms of utilities, attainable by the players involved, along with a disagreement outcome.

Suppose that $\mathcal{M} = \{1, 2, \cdots, M\}$ be the set of players of the bargaining game. They will either come to an agreement or fail to reach one. Assume a vector $\boldsymbol{x} = \{x_1, \cdots, x_M\}$ denoting all feasible allocation points, i.e. allocation of resources, for example: bandwidth, power etc., and a set of disagreement points $\boldsymbol{x}^0 = \{x_1^0, \cdots, x_M^0\}$. Note that these disagreement points denote the outcome for each player which is guaranteed in the absence of an agreement. $U_m(x_m)$ is the utility of player m, which is achieved if x_m is allocated via the bargaining process and $U_m^0(x_m^0)$ represents the utility of the same player when there is no agreement possible. Let $\mathcal{U}(\boldsymbol{x}) = \{(U_1(x_1), \cdots, U_M(x_M)) : (x_1, \cdots, x_M) \in \boldsymbol{x}\} \subset \mathbb{R}^M$ be the set of all possible utilities that the players can achieve and $\mathcal{U}^0(\boldsymbol{x}^0) = \{(U_1^0(x_1^0), \cdots, U_M^0(x_M^0) : (x_1^0, \cdots, x_M^0) \in \boldsymbol{x}^0\} \subset \mathbb{R}^M$ be the minimum desired utilities which each player gets without cooperation. The bargaining game can formally be defined as follows [81], [82].

Definition 2.5.1. *The pair* $(\mathcal{U}(\boldsymbol{x}), \mathcal{U}^0(\boldsymbol{x}^0))$ *defines the bargaining game where*

1. $\mathcal{U}(\boldsymbol{x})$ *is nonempty, compact (i.e., closed and bounded), and convex*

2. $\mathcal{U}^0(\boldsymbol{x}^0) \in \mathcal{U}(\boldsymbol{x})$,

3. $\max\{x_m : x \in \mathcal{U}(\boldsymbol{x})\}$ *exists* $\forall m \in \mathcal{M}$.

The set of all bargaining problems of the bargaining game defined as $(\mathcal{U}(\boldsymbol{x}),\mathcal{U}^0(\boldsymbol{x}^0))$ satisfying the aforementioned requirements are denoted by $\mathcal{B}$, also called the bargaining set, which includes all individually rational, Pareto-optimal payoff pairs[2].

The next step is to find the solution of the bargaining game. The axiomatic approach proposed by Nash gives a unique and fair Pareto-optimal solution [83], called the Nash bargaining solution (NBS) of a bargaining game. The NBS of a bargaining problem is defined as follows [81], [84]:

Definition 2.5.2. *Let* Γ *be a function and* $\Gamma\colon(\mathcal{U}(\boldsymbol{x}),\mathcal{U}^0(\boldsymbol{x}^0)) \to \mathbb{R}^M$. *A bargaining solution is then defined by the function* $\Gamma\colon\mathcal{B} \to \mathbb{R}^M$. $\mathcal{U}^\star$ *is said to be an NBS in* $\mathcal{U}(\boldsymbol{x})$ *for* $\mathcal{U}^0(\boldsymbol{x}^0)$, *if the following axioms are satisfied.*

A1 *Individual Rationality*[3]: $(U_m(x_m))^\star \geq (U_m^0(x_m^0))^\star$ for all m.

A2 *Feasibility*: $\mathcal{U}^\star \in \mathcal{U}(\boldsymbol{x})$.

A3 *Pareto-optimality*: The uniquely optimal solution $\mathcal{U}^\star$ is also Pareto-optimal.

A4 *Independence of Irrelevant Alternative*: If $\mathcal{U}^\star \in (\mathcal{U}(\boldsymbol{x}))' \subset \mathcal{U}(\boldsymbol{x})$ and $\mathcal{U}^\star = \Gamma(\mathcal{U}(\boldsymbol{x}),\mathcal{U}^0(\boldsymbol{x}^0))$, then $\mathcal{U}^\star = \Gamma((\mathcal{U}(\boldsymbol{x}))',\mathcal{U}^0(\boldsymbol{x}^0))$

A5 *Independence of Linear Transformation*: For any linear scale transformation ϕ, $\phi(\Gamma(\mathcal{U}(\boldsymbol{x}),\mathcal{U}^0(\boldsymbol{x}^0)) = \Gamma(\phi(\mathcal{U}(\boldsymbol{x})),\phi(\mathcal{U}^0(\boldsymbol{x}^0)))$.

[2] An element x of a vector $\boldsymbol{x}$ such that $x \in \boldsymbol{x}$ is said to be *Pareto-efficient* if there is no other element x' in $\boldsymbol{x}$ that all players like more than x. A Pareto-efficient point is sometimes Pareto-optimal and can not be improved on.

[3] An agreement is said to be individually rational if each player gets a utility under the agreement which is at least as large as a player can obtain without an agreement.

A6 *Symmetry*: If $\mathcal{U}^0(\boldsymbol{x}^0)$ is invariant under all exchanges of players such that

$(U_m(x_m), U'_m(x'_m)) \in \mathcal{U}(\boldsymbol{x}) \Leftrightarrow (U'_m(x'_m), U_m(x_m)) \in \mathcal{U}(\boldsymbol{x})$,

then $\Gamma_m(\mathcal{U}(\boldsymbol{x}, \mathcal{U}^0(\boldsymbol{x}^0)) = \Gamma'_m(\mathcal{U}(\boldsymbol{x}), \mathcal{U}^0(\boldsymbol{x}^0)) \forall m, m'$.

The axioms A1-A3 define how the NBS is located in the bargaining set $\mathcal{B}$. The axiom A4 states that eliminating feasible solutions should not affect the NBS outcome since it would not have been chosen by the players. This axiom indeed exhibits a unique phenomenon, and applying it to a particular resource-allocation problem leads to an optimal solution [85]. The axiom A5 implies that the bargaining solution is invariant to linear transformation of the utilities and disagreement point. The axiom A6 explains that players will converge to the same NBS if their utilities and disagreement points are the same. The fairness of an NBS is defined by the axioms A4-A6. The above six axioms lead to a uniquely optimal NBS called the generalised NBS and mathematically expressed by the following Theorem 2.5.1.

Theorem 2.5.1. *Genralized Nash Bargaining Solution (GNBS): A unique and fair NBS can be obtained for any user m by maximising the following Nash product term:*

$$\max_{\boldsymbol{x}} \quad \prod_{m \in \mathcal{M}} \Big(U_m(x_m) - U_m^0(x_m^0) \Big), \tag{2.1a}$$

$$\text{s.t.} \quad \Big\{ \Big(U_1(x_1), ..., (U_M(x_M) \Big) \Big\} \in \mathcal{U}(\boldsymbol{x}), \tag{2.1b}$$

$$U_m(x_m) > U_m^0(x_m^0) \forall m \tag{2.1c}$$

The NBS expressed by (2.1a)-(2.1c) can be transformed equivalently into the following optimisation problem [86], [87] and thus the following theorem can be obtained which is an extension of Theorem 2.5.1:

Theorem 2.5.2. *Let $U_m(\cdot) : \boldsymbol{x} \longrightarrow \mathbb{R}, m \in \mathcal{M}$ be a concave and upper-bounded function defined on $\boldsymbol{x}$, which is convex and compact and $\boldsymbol{x} \subset \mathbb{R}^M$. Now let, $G : \boldsymbol{x} \longrightarrow \mathbb{R}_+^\star$ be*

concave. Then $H = \log(G(\cdot)) : \mathbb{R}_+ \longrightarrow \mathbb{R}$ *is concave*[4]*. If* G *is injective then* H *is strictly concave. Further, let* $U_m(x_m); m \in \mathcal{M}$ *be injective on* $\boldsymbol{x}^0$*; then the equivalent GNBS can be expressed by the following optimisation:*

$$\max_{\boldsymbol{x}} \quad \sum_{m \in \mathcal{M}} \log\Big(U_m(x_m) - U_m^0(x_m^0)\Big), \tag{2.2a}$$

$$\text{s.t.} \quad \Big\{\Big(U_1(x_1), ..., (U_M(x_M)\Big)\Big\} \in \mathcal{U}(\boldsymbol{x}), \tag{2.2b}$$

$$U_m(x_m) > U_m^0(x_m^0) \forall m \tag{2.2c}$$

The equivalent GNBS expressed by (2.2a)-(2.2c) has a unique property. It states that the optimal allocation is the maximisation of a logarithmic objective function and thus can be utilised to preserve convexity and hence a unique, fair and optimal solution can be obtained. Furthermore, it has been proved that the maximisation of the sum of the logarithmic utility functions results in an allocation which is proportionally fair [88]. Thus the proportional fairness is an special case of GNBS.

2.5.2 Application of NBG in 5G HetNets

Fairness and Pareto-optimality are most exciting and desirable features in wireless networks consisting of a large number of nodes. The solution of an NBG achieves both. Inspired by its success in solving economics problems, an NBG was first proposed in [89] in the context of telecommunication for flow control. Bargaining games have been used ever since to allocate resources, to share resources, to solve offloading problems and so on. The authors of [90] propose an NBG-based framework to offload users and determine trade-offs between various efficiency parameters (such as EE and SE) and fairness. [86] proposes a bargaining game framework to characterise the bandwidth allocation and pricing policy of a wireless network which takes into account a user budget in a fair way

[4]All logarithms are to base 10 unless stated otherwise.

while maximising the network revenue. In [91] an optimal power-control problem has been studied from a cooperative-game perspective. Secondary users increase their SINR under an interference-protection constraint. In [85], a network-resource-sharing problem is solved using an NBG for multiuser multimedia applications. [92] presents two different bargaining models, namely sequential bargaining and concurrent bargaining. In a sequential bargaining model, a user bargains with one AP owner at a time for association. On the other hand, a user bargains with multiple AP owners in a concurrent bargaining model. It has been shown that AP owners lose benefit in concurrent bargaining. To improve economic gain, AP owners form groups. Practically, a wireless network consists of a large number of nodes. Therefore the chances of unfair situations increase, where some nodes are starved of resource. This results in an inefficient resource allocation and will have a higher impact in a 5G wireless system since, in 5G, the number of nodes is higher than in 4G. Therefore fairness will become extremely important in the context of 5G systems. This motivates us to apply the NBG-concept in a content-caching scenario where our aim is to ensure fair association among users for content-downloading purposes.

2.5.3 Evolutionary Game Theory

In this section we first introduce the basic concept of evolutionary game theory (EGT) and an evolutionary stable strategy (ESS). Replicator dynamics is briefly discussed, and provides a deterministic way to calculate the evolutionary equilibrium.

An evolutionary game mathematically models the concept of evolution by natural selection. It considers a large population of players that randomly interact with each other based on a pre-programmed strategy. The population evolves in such a way that the frequency with which a particular decision is made can change over time in response to the decisions made by all individuals. Initially, the population may consist of players using different strategies. The payoff to an individual adopting a strategy is defined

as the fitness for that type in the current population. The proportion of a population programmed with a particular strategy will increase if they have a higher fitness than the rest of the population. Therefore, players are likely to switch to those strategies that give better payoffs and move away from those that give poor payoffs. In this manner, those strategies that give a lower payoff are out-competed.

Definition 2.5.3. *An evolutionary game can be defined as* $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \pi(S_i)_{S_i \in \mathcal{S}})$ *where* $\mathcal{N}$ *is the set of players which constitute the population in an evolutionary game,* $\mathcal{S}$ *is the set of all strategies and* $\pi(S_i)_{S_i \in \mathcal{S}}$ *is the payoff obtained by using strategy* S_i*.*

Evolutionary Stable Strategy (ESS)

The concept of an ESS arose from the context of players' behaviour where a large population of players who are repeatedly and randomly matched in pairs play a game. In fact, ESS was first presented by Maynard Smith and Price via a seminal paper [93] for a monomorphic population [94], [95]. Note that a population of an evolutionary game is called monomorphic if every player adopts the same strategy and polymorphic if a proportion of the population has a different strategy, i.e. more than one strategy is present. The underlying assumptions of an ESS therefore are: i) players choose their strategies from identical sets, ii) the payoff to a player choosing a particular strategy against an opponent choosing an alternative is the same regardless of the identities or characteristics of the players; and iii) players cannot make their strategy choices conditional on any characteristics. The definition of an ESS states that an ESS is a strategy such that if all the members of a population adopt it, then no mutant strategy could invade the population under the influence of natural selection. Let us consider a player using a strategy S against a player using strategy $\hat{S}$. In that case, the expected payoff for the player using strategy S is $\pi(S, \hat{S})$. The definition of an ESS for a monomorphic population can be expressed as

Definition 2.5.4. *A strategy S^* is an ESS if and only if for all, $S \neq S^*$,*

1. $\pi(S, S^*) \leq \pi(S^*, S^*)$,

2. *if* $\pi(S, S^*) = \pi(S^*, S^*)$ *then* $\pi(S, S) < \pi(S^*, S)$.

Condition 1 defines the equilibrium condition and states that S^* is the best response strategy to itself and is hence a Nash equilibrium (NE). On the other hand, condition 2 is interpreted as a stability condition. Suppose that the incumbents play S^* and mutants play S. Then both conditions ensure that as long as the fraction of mutants playing S is not too large, the average payoff to S will fall short of that to S^*. However, the above definition of an ESS is only valid for a monomorphic population, in which every player uses the same strategy. For a polymorphic population, the specific type of evolutionary dynamics [95] called replicator dynamics is used to capture the evolution of a particular strategy in an evolutionary game. In the next section we give an overview of replicator dynamics.

Replicator Dynamics

The replicator dynamics introduced by Taylor and Jonker [96] reflects the role of a selection mechanism, i.e. a strategy in an evolutionary process. It is expressed as a system of ordinary differential equations (ODE) which does not consider mutations at all. Consider a large population of individuals, where each individual is programmed to adopt any of the pure strategies from the pure strategy set, $\mathcal{S} = \{1, ..., S\}$. Suppose that an individual adopts a pure strategy i, such that $i \in \mathcal{S}$. The population share of a particular strategy denotes the proportion of individuals in the total population that have adopted that strategy. Therefore, the population share of strategy i is x_i. The vector $\mathbf{x} = [x_1 x_2 ... x_s]^\intercal$ is called the population state, which includes the population share of all the strategies, $x_i \in \mathbf{x}$, where $\sum_{i=1}^{|\mathbf{S}|} x_i = 1$ is satisfied. Individuals with fixed strategies randomly interact

with other individuals. Based on the context of the game, the population share of those strategies that a yield a higher payoff will increase in proportion, which imitates the process of natural selection. The rate of a particular strategy adaptation of a group in the population is given by the replicator dynamics, expressed in terms of ODEs [97] as:

$$\dot{x_i} = x_i[\pi_i - \bar{\pi}], \tag{2.3}$$

where π_i is the payoff of each individual choosing strategy i and $\bar{\pi}$ is the average payoff of the entire population, which is given by:

$$\bar{\pi} = \sum_{i=1}^{|\mathcal{S}|} \pi_i x_i \, . \tag{2.4}$$

Thus the replicator dynamics presented by (2.3) states that the population share of strategy i will increase if the payoff associated with strategy i is higher than the average payoff, (i.e. $\pi_i > \bar{\pi}$).

2.5.4 Application of EGT in 5G HetNets

EGT has been applied to analyse wireless network and communication issues such as delay, spectrum access and network selection. For example EGT has been applied to derive delayed evolutionary game dynamics [98] that capture the phenomenon of the time delays of wireless networks in terms of feedback delay, content-caching delay, propagation delay etc. [99] utilises the concept of EGT to derive the necessary conditions to update the classic coalitional game theory. Interacting users decide upon whether to form or not to form a coalition for dynamically accessing spectrum. [100] proposes a potential game-based opportunistic spectrum access for CR-based wireless networks and uses the EGT to show that the Nash equilibrium (NE) of the potential game is an ESS. In addition a distributed learning algorithm is proposed based on an EGT, which achieves evolutionary stability with incomplete network information. A hierarchical game structure is proposed

in [101], where users perform a suitable service-provider (SP) selection based on an EGT framework. On the other hand, SPs compete between themselves and decide on the price per unit resource. Interaction among multiple SPs is modelled via a Stackelberg game and a non-cooperative game. Users will be distributed among the providers based on a rational choice that involves a trade-off between the price and the quality of service. Motivated by the capability of the EGT to capture the essence of the network dynamics, this thesis provides an EGT-based framework, where users' behaviour under the proposed pricing strategy is analysed by means of the replicator dynamics, in Chapter 3.

2.6 Summary

This chapter gives a comprehensive review on the recent research status about promising topics such as offloading and content caching. The 5G standardisation activity and industry-based research indicate that the initial 5G deployment will focus on HetNet-based technologies such as offloading and content caching to provide service, while massive-MIMO and mm-wave will be embodied in 5G networks in future. This chapter also provides a brief understanding about two game-theory-based algorithms, namely NBG and EGT, which are used in the following chapters to enable offloading and content caching in 5G networks. Both algorithms are used in this thesis to devise intelligent and efficient resource-allocation mechanisms for 5G HetNets which increase the utilisation of SBSs in a HetNet.

Chapter 3

Evolutionary Game Theory-based Offloading in 5G HetNet

3.1 Chapter Introduction

This chapter presents a novel price-based offloading strategy to provide a necessary framework for macrocells to control their population share. Unlike existing works [49], [54], [56], [58], [64], [102] the proposed pricing scheme guarantees that a proportion of macrocell users will always be offloaded to the small-cell tier. Tuning certain parameters of the pricing function, a macrocell can control its population share to avoid congestion. The price function combines two different pricing strategies, activated in two different cases. A macrocell activates *linear pricing* if its population share is small. Each MU pays a price to the macrocell which is proportional to the average data rate of that macrocellular network. As the population share of the macrocell network increases, the average data rate per user decreases. Also, the average data rate of a macrocell decreases because of cross-tier interference suffered by MUs, due to the proximity to SBSs. Therefore, the payment of each MU decreases under those two circumstances. As a result, a macrocell

therefore generates less revenue when both the interference and the number of macro users are large. By analysing this phenomenon, a *threshold price* in the macrocell price function is introduced. As soon as the average data rate falls below a certain threshold, called the *rate-threshold*, the second price component is activated by a macrocell. We call this price component as the threshold price. The rate-threshold thus works as the *trigger* for threshold price activation. Once the threshold price is activated, each MU has to pay a higher price than the linear price to remain in the macrocellular network. This motivates MUs to join nearby SBSs. In this process, a portion of the macrocell population is offloaded to its serious interferer. By doing so they avoid paying a high cost for a lesser resource (data rate) and avoid cross-tier interference as well. The numerical value of the rate-threshold can be fine-tuned by a macrocell to have an optimum population share with a view to minimising cross-tier interference and congestion.

SBSs, on the other hand, can increase their individual utility by selling redundant resources to MUs. We assume that small-cells adopt existing CRE techniques to accommodate MUs. The small-cell pricing strategy includes an access price and interference compensation. The access price is a fixed price paid for small-cell association. SBSs increase their transmit power by a certain level to increase their coverage. This allows MUs to connect with their nearest SBS seamlessly. However, increasing the transmit power will cause cross-tier interference. Interference compensation thus acts as a penalty for users who choose a small-cell instead of a macrocell. Thus it restricts distant MUs from joining a small-cell because of the cost and data rate trade-off. The numerical value of interference compensation is proportional to the number of active MUs who choose small-cells for association.

EGT provides an analytic framework for understanding competitive decision making, where a large numbers of players with different strategies interact in a dynamic scenario. In our model, MUs are the players that interact with an MBS and SBSs for association.

Table 3.1: List of Symbol Notations

Symbol	Description
$\mathcal{N}$	Set of MUs
$\mathcal{S}$	Set of the strategies
C_M	Macrocell price function
C_S	Small-cell price function
C_l	Linear cost component of macrocell
C_τ	Threshold cost component of macrocell
C_a	Access cost component of small-cell
C_i	Interference cost component of small-cell
R_M	Total data rate of macrocell
R_S	Total data rate of small-cell network
R_τ	Threshold data rate of macrocell

In this thesis, this interaction is modelled with EGT. MUs adapt to their choices based on the payoff they receive and will eventually distribute across the two-tier network. In other words, they will select either an SBS or an MBS based on the strategy, whichever yields them the higher payoff. Numerical results are provided by applying different network settings to emphasise the role of the proposed pricing strategies in offloading MUs to small-cells in order to shape network dynamics.

3.2 System Model

This section provides the notation of some important parameters. After that the network model is elaborated, with necessary assumptions. Finally the proposed pricing function is explained in Section 3.2.3.

3.2.1 Notation Summary

Since there are many notations in this chapter, for readers' convenience they are summarised in Table 3.1.

3.2.2 Network Model and Assumptions

A long term evolution (LTE) cellular network is considered, comprising a single macrocell and a large number of densely deployed small-cells overlaid inside the macrocell coverage area as shown in Fig. 3.1. We denote the macrocell index as M and the set of SBSs in the HetNet as $\mathcal{B} = \{B_1, B_2, ..., B_K\}$, where B_k, $k \in \mathcal{K} = \{1, 2, ..., K\}$ represents the k-th SBS. $\mathcal{N} = \{1, 2, .., N\}$ denotes the set of MUs in the HetNet that can connect either with an SBS or the macrocell. We assume that small-cells are operated in a hybrid access manner in order to allow MUs to join them. Furthermore, small-cells are deployed in underlay spectrum access mode to ensure efficient spectrum reuse and provide good transmission performance due to favorable channel characteristics. Both the macrocell and SBSs are connected to the core network with wired backhaul links such as optical fiber. For a channel model, we assume a similar model adopted in [49] and [54] because of the nature of the utility function developed in Section 3.3 and to maintain tractability for finding a feasible solution. We assume an underlay spectrum access based LTE HetNet deployment. Therefore cross-tier interference exists between the small-cell tier and the macrocell tier. The total data rate offered by the macrocell is R_M, shared equally among MUs, and R_S is the data rate offered to MUs by each small-cell. We assume that small-cells coordinate their transmission and avoid co-tier interference. Therefore data rate variation occurs in both tiers for two reasons: i) when cross-tier interference is high, and ii) when a significant number of users associate with a base station, resulting in network congestion.

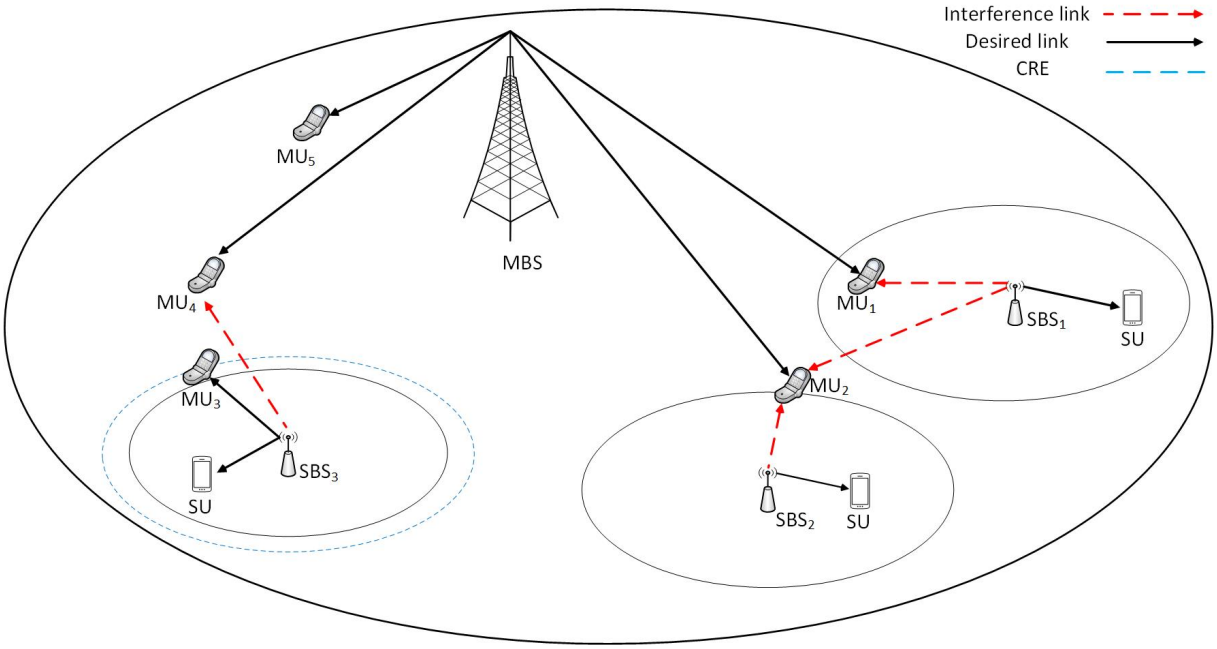


Figure 3.1: Demonstration of a two-tier dense HetNet.

Motivating example

The snapshot of our considered network model, given by Fig. 3.1, demonstrates the interference scenario in the HetNet. The existence of active SBSs near to MUs can result in detrimental interference, as an MU_1 is subjected to downlink cross-tier interference from SBS_1, while MU_2 perceives interference from both SBS_1 and SBS_2. The severity of the interference perceived by an MU on its downlink largely depends on the proximity to SBSs, i.e. the distance between an MU and SBSs. For example: SBS_2 is the most serious interferer for MU_2, though MU_2 perceives weak interference from SBS_1. Therefore, throughout the HetNet, MUs staying near to an SBS will suffer from interference, which in turn will reduce their received data rate. We assume that the macrocell has a limited bandwidth which is equally shared among MUs. Consequently, when the number of MUs in the macrocellular network increases, the data rate per user decreases. In a practical LTE cellular network the downlink received data rate of a user depends on the number

of assigned physical resource blocks (PRBs) and the received signal-to-interference-plus-noise-ratio (SINR). For example: if there are 5 users and 10 available PRBs, each user can have a maximum allocation of 2 PRBs. On the other hand, if there are 10 users, each user can get a maximum allocation of 1 PRB. Therefore the average data rate per user depends on the number of users present in the network and the instantaneous SINR. Thus interference and the number of MUs jointly affect the data rate of MUs when they are in the macrocellular network. SBSs, on the other hand, provide offloading opportunities for those MUs based on the assumption that SBSs have unused resource [35]. We assume that SBSs adopt an existing cell range expansion (CRE) technique such as power control to accommodate MUs. Such a case is demonstrated by the blue dotted line in Fig. 3.1, which indicates the expanded coverage area of an SBS. We assume that SBSs are allowed to increase their transmission power to a certain level to perform CRE. However, this results in interference to distant MUs as shown in Fig. 3.1. SBS_3 adjusts its transmission power to accommodate MU_3. Note that MU_4 is far enough to perceive cross-tier interference, but it is not close enough to be offloaded to SBSs. Taking these cases into consideration, both a macrocell and a small-cell pricing strategy is devised in this chapter.

3.2.3 Pricing Strategies

Pricing any resource can be viewed as a control mechanism. Here the goal of the pricing is to induce offloading, i.e. it works as a punishing strategy to achieve the goal of offloading. One of the most commonly used pricing strategies in practical wireless networks is fixed pricing, where users pay a fixed charge for given amount of resource [49], [102]. A pricing concept has also been adopted in [103] for developing an uplink power control using a non-cooperative game. A linear-pricing based model is used for uplink power control in a CDMA-based wireless network. In the linear pricing strategy [51], [54], [58] a user generally pays for per unit of resource and the total amount of payment is proportional

to the total amount of the resources consumed by the user. For example: the payment of a user to the network operator increases linearly as the total number of users increases in that network [54]. On the contrary, a novel approach has been taken to formulate the price function for the network resource, detailed as follows:

Macrocell Pricing Strategy

Pricing strategies can shape the tele-traffic dynamics of the wireless network. Unlike a linear-pricing-based model [103] used for CDMA uplink power control, we have combined a linear and a threshold pricing strategy to obtain the objective of macrocell-controlled offloading. Motivated by the potential of influencing MUs by applying the appropriate price theory, the following price function is proposed:

$$\begin{aligned} C_M &= C_\tau + C_l \\ &= C_f R_\tau + C_f(\bar{R} - R_\tau)^+, \end{aligned} \tag{3.1}$$

where C_M is the macrocell's price function. By definition x^+ is given by:

$$x^+ = \max(x, 0) = \begin{cases} x & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases} \tag{3.2}$$

C_M is comprised of two types of price components, namely a threshold price and a linear price. C_l represents the linear price component and is equivalent to $C_f(\bar{R} - R_\tau)^+$. $\bar{R}$ and R_τ represent the average macrocell data rate and the rate-threshold respectively. C_f is a fixed cost. C_τ is the threshold price component and is defined as $C_\tau = C_f R_\tau$. The graphical representation of the price function is given in Fig. 3.2.

Equation (3.1) yields two different cases:

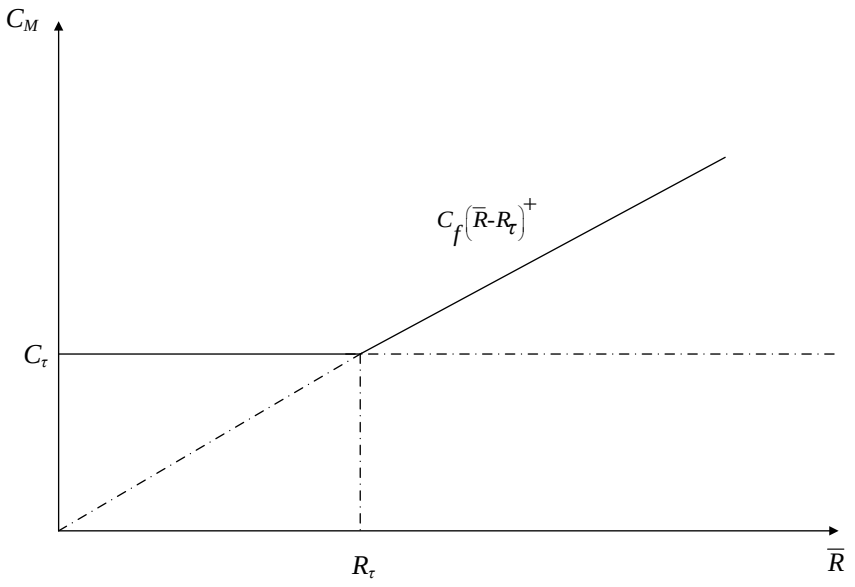


Figure 3.2: Graphical illustration of the price function.

Case 1: $(\bar{R} - R_\tau) > 0$

$$\begin{aligned} C_M &= C_\tau + C_f(\bar{R} - R_\tau) \\ &= C_f R_\tau + C_f \bar{R} - C_f R_\tau \\ &= C_f \frac{R_M}{|\mathcal{N}_\mathcal{M}|}, \end{aligned} \tag{3.3}$$

where $\bar{R} = \frac{R_M}{|\mathcal{N}_\mathcal{M}|}$, the value of which depends on the number of MUs connected with the macrocell. $|\mathcal{N}_\mathcal{M}|$ is the cardinality of set $\mathcal{N}_\mathcal{M}$, denoting the total number of MUs that select the macrocell for association. We see that the macrocell adopts a linear pricing strategy when $\bar{R} > R_\tau$. Each user willing to connect with a macrocell pays price $C_M = C_l$, which varies linearly with an increase or decrease in the average data rate.

Case 2: $(\bar{R} - R_\tau) < 0$

$$\begin{aligned} C_M &= C_\tau + 0 \\ &= C_f R_\tau. \end{aligned} \tag{3.4}$$

The macrocell thus adopts the threshold pricing strategy when $\bar{R} < R_\tau$. The numerical value of the threshold price component depends on the rate-threshold R_τ. The function of this rate-threshold is very crucial in the context of offloading, since it predominantly controls the macrocell population share. Higher values of R_τ ensure activation of the threshold price component for smaller numbers of MUs in the macrocell. On the other hand, by lowering R_τ a macrocell can entice more MUs to remain connected with the macrocell. Therefore, the macrocell can control its population share (PS) by changing R_τ. Note that the macrocell needs to keep track of its population share in order to make a decision about the rate-threshold.

Small-cell Pricing Strategy

The pricing strategy of a small-cell is given by:

$$C_S = C_a + C_i|\mathcal{N}_\mathcal{S}|, \tag{3.5}$$

where C_a, C_i are the access price and the interference price respectively. Both are a fixed charge. $|\mathcal{N}_\mathcal{S}|$ is the cardinality of set $\mathcal{N}_\mathcal{S}$, denoting the total number of MUs that select a small-cell for association. Therefore the total payment of a user selecting a small-cell depends on the number of users that choose the small-cell for association. The interference compensation serves two purposes. Firstly, it restricts a small-cell from being overcrowded. Secondly, it prevents a small-cell from increasing its transmission power abruptly to attract distant MUs. We assume that the macrocell feeds back the location information of the MUs via a common control channel to small-cells to operate the pricing strategy and perform CRE. Stackelberg game is generally used for computing the optimal

price in a hierarchical pricing structure in [55] for CRE-based cell association, in [56] for determining the bandwidth price for macrocell and small-cell association, and in [57] for determining the spectrum leasing price where a macrocell leases a part of its spectrum to small-cells to accommodate offloaded MUs. However our aim is to analytically determine the evolution of population and justify the macrocell-controlled offloading in the HetNet. Therefore an evolutionary game-theoretic framework is more suitable [49], [52], [54], [55], and it is introduced to capture the network dynamics of a large population in the following section.

3.3 Network Selection Game: Modelling The Evolution of MUs

Before detailing the network selection game we briefly discuss the strategic decision making of players in an evolutionary game. One of the most important assumptions of EGT is that players playing an evolutionary game have bounded rationality[1]. The concept of bounded rationality is neither irrationality nor full rationality[2]. Unlike rational entities, entities with bounded rationality exhibit non-optimising behaviour and they are guided by an aspiration of adaptation [107]. Thus in the context of rationality an evolutionary game provides a solution, where players slowly change their strategy to eventually reach that solution. On the other hand, a classical single-player non-cooperative game assumes full rationality and players make choices such that they reach immediately to the desired

[1]Bounded rationality is a term proposed by H. A. Simon [104], which explains a strategic decision-making procedure among individuals, where the rationality of individuals is limited by the information they have, the cognitive limitations of their minds and the finite amount of time they have to make a decision [105], [106], [107].

[2]The term full rationality or rationality considers individuals as Bayesian maximisers of subjective utility [107]. Full rationality requires unlimited cognitive capabilities.

solution [49], [54].

We model the network-selection procedure of players in the form of an evolutionary game. Boundedly rational players involved in the network-selection game will compete for the better data rate provider by selecting a suitable base station. Initially, a player may choose a base station randomly. However the user will adapt its network-selection decision periodically by observing the average received data rate variation of a macrocell and small-cells along with pricing strategies of a macrocell and small-cells to gain a better payoff. Thus all players participating in the evolutionary game-theoretic network-selection procedure will evolve over time depending on the received average data rate and the pricing strategies of macocells and small-cells. The following section presents the framework of the evolutionary game for network selection with our proposed pricing strategies.

3.3.1 Essence of The Network-Selection Game

The essence of the network selection-game is as follows:

Players: The large population of MUs, i.e. the players of the game are defined by the set $\mathcal{N} = \{1, 2, ..., N\}$.

Set of Strategies: The MUs will select either the macrocell or any of the small-cells, whichever yields them the highest payoff under the pricing strategies of both networks. We define the set of strategies for each type of player as $\mathcal{S} = \{S_M, S_S\}$.

Population Share (PS): The set of MUs constitutes the population in an evolutionary game. The players will adopt one of the two strategies and they will be distributed across the two-tier network accordingly. We assume that a centralised controller, for example a macrocell, can collect and relay required information (the price and the population share of each tier) from/to every MU through a common control channel [54]. A proportion of the MUs will select the macrocell, i.e., adopt strategy S_M for association. Therefore the

population share (PS) of MUs for macrocell selection is defined as:

$$X_M = |\mathcal{N}_\mathcal{M}|/|\mathcal{N}|, \tag{3.6}$$

where $|\mathcal{N}_\mathcal{M}|$ is the cardinality of set $\mathcal{N}_\mathcal{M}$; $\mathcal{N}_\mathcal{M} \subset \mathcal{N}$. Similarly the PS of MUs who opt to select small-cells for association is defined as:

$$X_S = |\mathcal{N}_\mathcal{S}|/|\mathcal{N}|, \tag{3.7}$$

where $|\mathcal{N}_\mathcal{S}|$ is the cardinality of set $\mathcal{N}_\mathcal{S}$; $\mathcal{N}_\mathcal{S} \subset \mathcal{N}$, which denotes the number of MUs willing to connect with a small-cell. Here $X_M + X_S = 1$.

Payoff Function: According to EGT, the payoff function or fitness of a player (MU) quantifies the amount of profit it could gain by adopting a particular strategy [96], [97]. We use the following non-increasing function in the interval of $X_M \in [0, 1]$ to obtain the payoff of an MU (using strategy S_M) for selecting a macrocell:

$$\pi_M(X_M) = \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_M, \tag{3.8}$$

where the first component of the payoff function is called the utility which is achieved by a MU for selecting a macrocell for association, in which ω_M is the weighting factor for macrocell selection and R_M is the total data rate provided by the macrocell. The second term, C_M, is the cost incurred by an MU for receiving the service of the macrocell.

Here we adopt a homogeneous utility function to model the small-cell selection for the sake of reaching a tractable solution. Under the homogeneous utility model all small-cells work as a single entity [54], [102] and [108] and they charge same price to each MU for small-cell association. Note that each small-cell can have different pricing schemes [49] prompting MUs to consider multiple cell selection strategies. However evaluating such a model will result in a higher computational complexity and hence convergence will be slower. We use the following non-increasing function in the interval of $X_S \in [0, 1]$ to

quantify the payoff of an MU (using strategy S_s) for selecting a small-cell:

$$\pi_S(X_S) = \frac{\omega_S R_S}{|\mathcal{N}|X_S} - C_S, \tag{3.9}$$

where ω_S is the weighting factor of small-cells, R_S is the total data rate provided by the small-cell and C_S is the cost incurred by an MU for receiving the service of the small-cell. The payoff function of an MU selecting a small-cell can be written as a function of the macrocell PS (X_M). Note that the payoff function now becomes a non-decreasing function in the interval of $X_M \in [0, 1]$ and, after applying (3.5), becomes:

$$\pi_S(X_M) = \frac{\omega_S R_S}{|\mathcal{N}|(1 - X_M)} - C_a - C_i|\mathcal{N}|(1 - X_M), \tag{3.10}$$

where $|\mathcal{N}_\mathcal{S}| = |\mathcal{N}|(1 - X_M)$.

3.3.2 Network Selection Dynamics and Equilibrium Computation

In this section, the replicator dynamics to compute the evolutionary equilibrium is derived, which reflects how MUs will distribute across the two-tier network after strategic interaction with the macrocell and small-cells. In a dynamic evolutionary game, each player of a population is called a replicator and each replicator is able to replicate itself through the process of imitation, mutation and selection. The process of replication takes place over time and is modelled by using a set of ODEs called replicator dynamics [49], [54]. Replicator dynamics is a very important tool in the context of an evolutionary game since it provides a deterministic way to compute the population share, i.e. the proportion of players who choose different strategies at a particular point in time. Recently, replicator dynamics is used to solve practical wireless networking problems. In [109], replicator dynamics is used to solve an EGT-based resource-allocation problem in orthogonal frequency-division multiple access (OFDMA)-based small-cell networks. In [54],

replicator dynamics is used to determine the evolution of population in IEEE 80.216, CDMA cellular and IEEE 802.11 wireless networks. Likewise we use replicator dynamics for our wireless networking problem to capture the interplay of MUs with the macrocell and small-cell network by determining the parameters X_M, X_S for the two different cases, i.e. for case 1 and case 2 respectively, by utilising replicator dynamics.These parameters give the proportion of MUs that choose a macrocell or a small-cell, respectively, under the proposed pricing functions. Using (2.3) the replicator dynamics [97] of the two-tier HetNet under the proposed pricing strategy can be written as:

$$\begin{aligned}\dot{X}_M &= \sigma X_M[\pi_M - \bar{\pi}] \\ &= \sigma X_M[\pi_M - X_M\pi_M - X_S\pi_S] \\ &= \sigma X_M[\pi_M - X_M\pi_M - (1 - X_M)\pi_S] \\ &= \sigma X_M(1 - X_M)[\pi_M - \pi_S],\end{aligned} \tag{3.11}$$

where σ is the evolutionary speed constant and explains how fast an evolutionary game will approach convergence, $\bar{\pi}$ is the average payoff and $\bar{\pi} = X_M\pi_M + X_S\pi_S$. We now analyse the network selection dynamics by applying the two types of price function.

Selection Dynamics in Case 1

Note that the macrocell price function is comprised of two different pricing strategies. Activation of each pricing strategy depends upon the number of MUs associated in a macrocellular network. Inserting (3.3) into (3.11) gives the payoff of a MU for macrocell association in case 1 as:

$$\pi_{M,1}(X_M) = \frac{\omega_M R_M}{|\mathcal{N}|X_M} - \frac{C_f R_M}{|\mathcal{N}|X_M}. \tag{3.12}$$

The replicator dynamics can be written as:

$$\begin{aligned}\dot{X}_M &= \sigma X_M(1-X_M)[\pi_{M,1}(X_M) - \pi_S(X_M)] \\ &= \sigma X_M(1-X_M)\left[\frac{\omega_M R_M}{|\mathcal{N}|X_M} - \frac{C_f R_M}{|\mathcal{N}|X_M} - \frac{\omega_S R_S}{|\mathcal{N}|(1-X_M)} + C_a + C_i|\mathcal{N}|(1-X_M)\right].\end{aligned} \tag{3.13}$$

The equilibrium point of an evolutionary game can be obtained by setting the left-hand side of the replicator dynamics (3.13) to zero, i.e., $\dot{X}_M = 0$. The solutions $X_M^* = 0$ and $X_M^* = 1$ are called trivial fixed points, also known as boundary evolutionary equilibria of the evolutionary game. The boundary evolutionary equilibria are not stable, since any small perturbation will cause the system to deviate from the equilibrium state [109]. Furthermore trivial fixed points do not provide a unique solution. Therefore we need to find non-trivial fixed points which will give a unique solution of the evolutionary game if and only if it is an interior evolutionary equilibrium. Non-trivial fixed points are given by the following equation:

$$\left[\frac{\omega_M R_M}{|\mathcal{N}|X_M} - \frac{C_f R_M}{|\mathcal{N}|X_M} - \frac{\omega_S R_S}{|\mathcal{N}|(1-X_M)} + C_a + C_i\mathcal{N}(1-X_M)\right] = 0. \tag{3.14}$$

Equation (3.14) finally takes the following form:

$$X_M^3 C_N|\mathcal{N}|^2 - X_M^2(C_a|\mathcal{N}| + 2C_i|\mathcal{N}|^2) - X_M(\omega_M R_M - C_f R_M + \omega_S R_S - C_a|\mathcal{N}| - 2C_i|\mathcal{N}|^2) + \omega_M R_M - C_f R_M = 0\,. \tag{3.15}$$

Equation (3.15) is a cubic and has 3 roots, i.e. the replicator dynamics has three more fixed points. Note that any fixed point X_M^* of the replicator dynamics is feasible if it is an interior evolutionary equilibrium [109]. The interior evolutionary equilibrium x_i is defined by the case when $x_i \in (0,1), \forall i \in \mathcal{S}$. In order to find the closed form expression of the feasible fixed points, the cubic needs to be solved. However, solving the cubic algebraically is not insightful. Instead, it has been shown that a feasible fixed point exists [109] by providing a necessary proof using the existence Theorem 3.4.1 in Section 3.4.1.

Selection Dynamics in Case 2

When the threshold price is active, i.e., in case 2, the payoff of a MU is found by substituting equation (3.4) in equation (3.11):

$$\pi_{M,2}(X_M) = \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f R_\tau. \tag{3.16}$$

From (3.11), the network selection dynamics can be derived as follows when the threshold price is active:

$$\begin{aligned}\dot{X}_M =& \sigma X_M(1-X_M)[\pi_{M,2}(X_M) - \pi_S(X_M)] \\ &= \sigma X_M(1-X_M)\left[\frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f R_\tau - \frac{\omega_S R_S}{|\mathcal{N}|(1-X_M)} + C_a + C_i|\mathcal{N}|(1-X_M)\right].\end{aligned} \tag{3.17}$$

Similarly, by setting $\dot{X}_M = 0$, two trivial fixed points can be found, $X_M^* = 0$ and $X_M^* = 1$, as well as the following cubic equation:

$$\begin{aligned}&X_M^3 C_i|\mathcal{N}|^2 + X_M^2(C_f R_\tau|\mathcal{N}| - C_a|\mathcal{N}| - 2C_i|\mathcal{N}|^2) - \\ &\qquad X_M(\omega_M R_M + C_f R_\tau|\mathcal{N}| + \omega_S R_S - C_a|\mathcal{N}| - C_i|\mathcal{N}|^2) + \omega_M R_M = 0.\end{aligned} \tag{3.18}$$

By solving (3.18), three more fixed points can be obtained for case 2. However they must each satisfy the condition of interior evolutionary equilibrium to be considered as a feasible solution. For the reason outlined in Case 1, our aim is to prove that a solution indeed exists, which is detailed in the following section by rigorously analysing the payoff functions.

3.4 Theoretical Analysis

3.4.1 Existence of Evolutionary Equilibrium

The concepts of existence and uniqueness are applied in order to prove that the equilibrium points indeed exist. Recall the general equation (3.8), which describes the payoff for each user connecting to a macrocell. Substitute the value of C_M in (3.8) to give:

$$\begin{aligned}
\pi_M(X_M, \mathcal{X}) &= \frac{\omega_M R_M}{|\mathcal{N}|X_M} - [C_f R_\tau + C_f(\bar{R} - R_\tau)^+] \\
&= \frac{\omega_M R_M}{|\mathcal{N}|X_M} - \left[C_f R_\tau + C_f\left(\frac{R_M}{|\mathcal{N}|X_M} - R_\tau\right)^+\right] \\
&= \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f\left[R_\tau + \max\left(0, \frac{R_M}{|\mathcal{N}|X_M} - R_\tau\right)\right] \\
&= \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f\left[\max\left(R_\tau, R_\tau + \frac{R_M}{|\mathcal{N}|X_M} - R_\tau\right)\right] \\
&= \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f\left[\max\left(R_\tau, \frac{R_M}{|\mathcal{N}|X_M}\right)\right] \\
&= \frac{\omega_M R_M}{|\mathcal{N}|X_M} + \left[\min\left(-C_f R_\tau, -C_f \frac{R_M}{|\mathcal{N}|X_M}\right)\right] \\
&= \min\left(\frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f R_\tau, \frac{\omega_M R_M}{|\mathcal{N}|X_M} - C_f \frac{R_M}{|\mathcal{N}|X_M}\right) \\
&= \min(\pi_{M,2}(X_M), \pi_{M,1}(X_M)).
\end{aligned} \tag{3.19}$$

Therefore we have to find the solution of the evolutionary game defined as:

$$\mathcal{G} = \langle \mathcal{N}, \mathcal{X}, \boldsymbol{\pi} \rangle, \tag{3.20}$$

where $\boldsymbol{\pi} = \{\pi_M(X_M), \pi_S(X_M)\}$ is the set of payoff functions and $\pi_M(X_M) = \min(\pi_{M,2}(X_M), \pi_{M,1}(X_M))$. The replicator dynamics can thus be expressed as:

$$\dot{X}_M = \sigma X_M(1 - X_M)\Delta(X_M), \tag{3.21}$$

where $\Delta(X_M) = \pi_M(X_M) - \pi_S(X_M)$, is a function of X_M, that defines the difference between the macrocell and small-cell payoffs. At equilibrium, $\dot{X}_M = 0$, setting $\dot{X}_M = 0$

yields $\Delta(X_M) = 0$ and also gives 0 and 1 as fixed points of the replicator dynamics. Now let us concentrate on the following equation:

$$\begin{aligned} &\Delta(X_M) = 0 \\ &\Rightarrow \pi_M(X_M) - \pi_S(X_M) = 0 \\ &\Rightarrow \min(\pi_{M,2}(X_M), \pi_{M,1}(X_M)) - \pi_S(X_M) = 0\,. \end{aligned} \tag{3.22}$$

Suppose that X_M^* is a fixed point of the replicator dynamics (3.21), achieved by solving equation (3.22) such that $\Delta(X_M^*) = 0$. If X_M^* is an interior evolutionary equilibrium, it will eventually be a feasible solution of evolutionary game $\mathcal{G}$. Thus X_M^* will be an interior evolutionary equilibrium if and only if $X_M^* \in (0, 1)$. The following theorem can be obtained applying the intermediate value theorem for continuous functions.

Theorem 3.4.1. *Suppose that $\pi_M(X_M)$ and $\pi_S(X_M)$ are both real-valued continuous functions whose domain is $[\alpha, \beta]$, also α and β are two real numbers such that $\alpha < \beta$. Suppose that $\pi_M(X_M)$ and $\pi_S(X_M)$ are decreasing and increasing respectively. Suppose that the range of $\pi_M(X_M)$ and $\pi_S(X_M)$ are the entire closed interval $[\pi_M(\alpha), \pi_M(\beta)]$ and $[\pi_S(\alpha), \pi_S(\beta)]$ respectively. If X_M^* is any point between $[\alpha, \beta]$ such that $\pi^* := \pi_M(X_M^*)$ and $\pi^* := \pi_S(X_M^*)$, then X_M^* is a solution of the replicator dynamics expressed by (3.21).*

Proof. To prove the existence of an evolutionary equilibrium by applying the intermediate value theorem let us consider μ and ϵ as two variables. Both μ and ϵ are defined in terms of very small positive real number and $\mu \neq \epsilon$. Consider μ is large enough and ϵ is small enough such that the following inequalities can be written:

$$\pi_M(\epsilon) > \mu, \tag{3.23}$$

and

$$\pi_S(1 - \epsilon) > \mu. \tag{3.24}$$

Recalling that both payoff functions $\pi_M(X_M)$ and $\pi_S(X_M)$ are bounded between 0 and 1, both (3.23) and (3.24) follow because they are continuous and approach to infinity at 0 and 1 respectively. Therefore the following inequalities are true based on the above assumptions:

$$\pi_S(\epsilon) < \pi_S(0) + \delta, \tag{3.25}$$

and

$$\pi_M(1-\epsilon) < \pi_M(1) + \delta. \tag{3.26}$$

(3.25) and (3.26) follow continuity. Subtracting (3.23) from (3.25) gives:

$$\pi_S(\epsilon) - \pi_M(\epsilon) < \pi_S(0) + \delta - \mu. \tag{3.27}$$

Also subtracting (3.24) from (3.26) gives:

$$\pi_S(1-\epsilon) - \pi_M(1-\epsilon) < \mu - \pi_M(1) - \delta. \tag{3.28}$$

Finally,

$$\mu > \max[(\pi_S(0) + \delta), (\pi_M(1) + \delta)]. \tag{3.29}$$

Therefore, there exists a fixed point $X_M^*: \pi_S(X_M^*) - \pi_M(X_M^*) = 0$. This completes the proof. ■

3.4.2 Uniqueness of Evolutionary Equilibrium

To show uniqueness, consider any point X_M^0, such that $X_M^0 \in (0,1)$ and $X_M^0 \neq X_M^*$. Also assume that $\pi^0 := \pi_M(X_M^0) := \pi_S(X_M^0)$ at $X_M = X_M^0$. If $X_M \neq X_M^*$ and $X_M > X_M^*$, we can write the following inequality:

$$\begin{aligned} \pi_M(X_M) - \pi_S(X_M) &< \pi_M^*(X_M^*) - \pi_S^*(X_M^*) \\ &\Rightarrow \pi_M(X_M) - \pi_S(X_M) < 0\,. \end{aligned} \tag{3.30}$$

Also when $X_M < X_M^*$, there is the following inequality:

$$\begin{aligned} \pi_M(X_M) - \pi_S(X_M) &> \pi_M^*(X_M^*) - \pi_S^*(X_M^*) \\ \Rightarrow \pi_M(X_M) - \pi_S(X_M) &> 0\,. \end{aligned} \tag{3.31}$$

The above inequalities, expressed by (3.30) and (3.31), guarantee that X_M^* is the unique solution in the domain of $X_M \in (0, 1)$.

3.4.3 Stability of Evolutionary Equilibrium

This section analyses the stability of the solution of the replicator dynamics (3.21). It will be shown that the evolutionary equilibrium is asymptotically stable by applying Lyapunov's direct method. First define the Lyapunov function $\mathcal{L}(X_M)$ for the replicator dynamics expressed by (3.21) as:

$$\mathcal{L}(X_M) = (X_M - X_M^*)^2\,. \tag{3.32}$$

Now the following theorem provides necessary conditions for the replicator dynamics $\dot{X}_M$ to have a stable point X_M^* at the equilibrium.

Theorem 3.4.2. *Let $\mathcal{L}: \mathcal{O} \to \mathbb{R}$ be a differentiable function defined on an open set $\mathcal{O}$ in $\mathbb{R}^n$, which contains the equilibrium point X_M^* of the replicator dynamics $\dot{X}_M$. Suppose further that*

(a) *$\mathcal{L}(X_M^*) = 0$ and $\mathcal{L}(X_M) > 0$ if $X_M \neq X_M^*$;*

(b) *If $\dot{\mathcal{L}}(X_M) \leq 0$ in $\mathcal{O} - X_M^*$, X_M^* is stable.*

(c) *Furthermore, if $\dot{\mathcal{L}}(X_M) < 0$ in $\mathcal{O} - X_M^*$, X_M^* is asymptotically stable.*

Proof. The function $\mathcal{L}(X_M)$, if it satisfies (a) and (b), is a Lyapunov function for X_M^*. Furthermore, $\mathcal{L}(X_M)$ is a strict Lyapunov function [110] if it satisfies (c). At $X_M = X_M^*$,

we see that $\mathcal{L}(X_M^*) = 0$. Also $\mathcal{L}(X_M) > 0$ for all X_M, except at $X_M \neq X_M^*$. Therefore, the condition (a) holds for (3.32). Now take the derivative of (3.32) with respect to t to obtain:

$$\dot{\mathcal{L}}(X_M) = 2(X_M - X_M^*)\left[\frac{d(X_M)}{dt} - \frac{d(X_M^*)}{dt}\right]. \tag{3.33}$$

Note that $\frac{d(X_M^*)}{dt} = 0$ at the equilibrium point. By substituting $\frac{d(X_M)}{dt}$, i.e., $\dot{X}_M$, in the right hand side of (3.33), we get:

$$\begin{aligned}
\dot{\mathcal{L}}(X_M) &= 2\sigma X_M(1 - X_M)(X_M - X_M^*)[\pi_M(X_M) - \pi_S(X_M)] \\
&= A(X_M - X_M^*) \\
&\quad [\min(\pi_{M,2}(X_M), \pi_{M,1}(X_M)) - \pi_S(X_M)] \\
&= A(X_M - X_M^*) \\
&\quad [\min((\pi_{M,2}(X_M) - \pi_S(X_M)), (\pi_{M,1}(X_M) - \pi_S(X_M)))]\,,
\end{aligned} \tag{3.34}$$

where $A = 2\sigma X_M(1 - X_M)$ is an arbitrary constant and $A > 0$. In order to prove $\dot{\mathcal{L}}(X_M) < 0$, the following two conditions must be true. First, when $X_M - X_M^* > 0$

$$[\min((\pi_{M,2}(X_M) - \pi_S(X_M)), (\pi_{M,1}(X_M) - \pi_S(X_M)))] < 0\,. \tag{3.35}$$

Second, when $X_M - X_M^* < 0$

$$[\min((\pi_{M,2}(X_M) - \pi_S(X_M)), (\pi_{M,1}(X_M) - \pi_S(X_M)))] > 0\,. \tag{3.36}$$

Note that at equilibrium $\pi_M^*(X_M^*) = \pi_S^*(X_M^*)$, therefore the following equality can be written:

$$\min(\pi_{M,2}^*(X_M), \pi_{M,1}^*(X_M)) = \pi_S^*(X_M^*)\,. \tag{3.37}$$

The following inequalities can be written when $X_M - X_M^* > 0$

$$\pi_{M,1}(X_M) < \pi_{M,1}^*(X_M), \tag{3.38}$$

$$\pi_{M,2}(X_M) < \pi^*_{M,2}(X_M), \tag{3.39}$$

and

$$\pi_S(X_M) > \pi^*_S(X_M) \Rightarrow -\pi_S(X_M) < -\pi^*_S(X_M)\,. \tag{3.40}$$

The following two inequalities are achieved by adding (3.40) to (3.38) and (3.39) respectively:

$$\pi_{M,1}(X_M) - \pi_S(X_M) < \pi^*_{M,1}(X_M) - \pi^*_S(X_M), \tag{3.41}$$

and

$$\pi_{M,2}(X_M) - \pi_S(X_M) < \pi^*_{M,2}(X_M) - \pi^*_S(X_M). \tag{3.42}$$

Suppose that a, b, c and d are non-zero real numbers such that $a < c$ and $b < d$. In that case, the following inequality condition is true:

$$\min(a, b) \leq \min(c, d)\,. \tag{3.43}$$

Applying the equation (3.43) to (3.41) and (3.42) gives:

$$\begin{aligned} &\min\Big[(\pi_{M,1}(X_M) - \pi_S(X_M)), (\pi_{M,2}(X_M) - \pi_S(X_M))\Big] \\ &< \min\Big[(\pi^*_{M,1}(X_M) - \pi^*_S(X_M)), (\pi^*_{M,2}(X_M) - \pi^*_S(X_M))\Big]. \end{aligned} \tag{3.44}$$

At equilibrium $\pi^*_{M,1}(X_M) - \pi^*_S(X_M) = 0$ and $\pi^*_{M,2}(X_M) - \pi^*_S(X_M) = 0$. Therefore inequality (3.44) takes the following form:

$$\min\Big[(\pi_{M,1}(X_M) - \pi_S(X_M)), (\pi_{M,2}(X_M) - \pi_S(X_M))\Big] < 0. \tag{3.45}$$

Therefore the inequality defined by (3.35) is true when $X_M - X^*_M > 0$. Now consider the condition for $X_M - X^*_M < 0$. When $X_M - X^*_M < 0$

$$\pi_{M,1}(X_M) > \pi^*_{M,1}(X_M), \tag{3.46}$$

$$\pi_{M,2}(X_M) > \pi^*_{M,2}(X_M), \tag{3.47}$$

and

$$\pi_S(X_M) < \pi^*_S(X_M) \Rightarrow -\pi_S(X_M) > -\pi^*_S(X_M). \tag{3.48}$$

Adding (3.48) to (3.46) and (3.47), respectively, gives the following inequalities:

$$\pi_{M,1}(X_M) - \pi_S(X_M) > \pi^*_{M,1}(X_M) - \pi^*_S(X_M), \tag{3.49}$$

and

$$\pi_{M,2}(X_M) - \pi_S(X_M) > \pi^*_{M,2}(X_M) - \pi^*_S(X_M)\,. \tag{3.50}$$

Now we can write the following inequality by applying (3.43) to the inequalities given by (3.49) and (3.50):

$$\begin{aligned} &\min\Big[(\pi_{M,1}(X_M) - \pi_S(X_M)), (\pi_{M,2}(X_M) - \pi_S(X_M))\Big] \\ &> \min\Big[(\pi^*_{M,1}(X_M) - \pi^*_S(X_M)), (\pi^*_{M,2}(X_M) - \pi^*_S(X_M))\Big]\,. \end{aligned} \tag{3.51}$$

By applying the equilibrium condition the following is obtained:

$$\min\Big[(\pi_{M,1}(X_M) - \pi_S(X_M)), (\pi_{M,2}(X_M) - \pi_S(X_M))\Big] > 0\,. \tag{3.52}$$

As a result, the inequality defined by equation (3.36) is satisfied. Thus it can be claimed that $\mathcal{L}(\dot{X}_M) < 0$ when $X_M - X^*_M > 0$ and $X_M - X^*_M < 0$. Therefore both conditions (b) and (c) hold and the solution of the replicator dynamics X^*_M is asymptotically stable. This completes the proof. ■

3.5 Analytical Result

This section analyses the numerical results obtained from solving the network selection dynamics under the proposed pricing strategies. We observe how different parameters

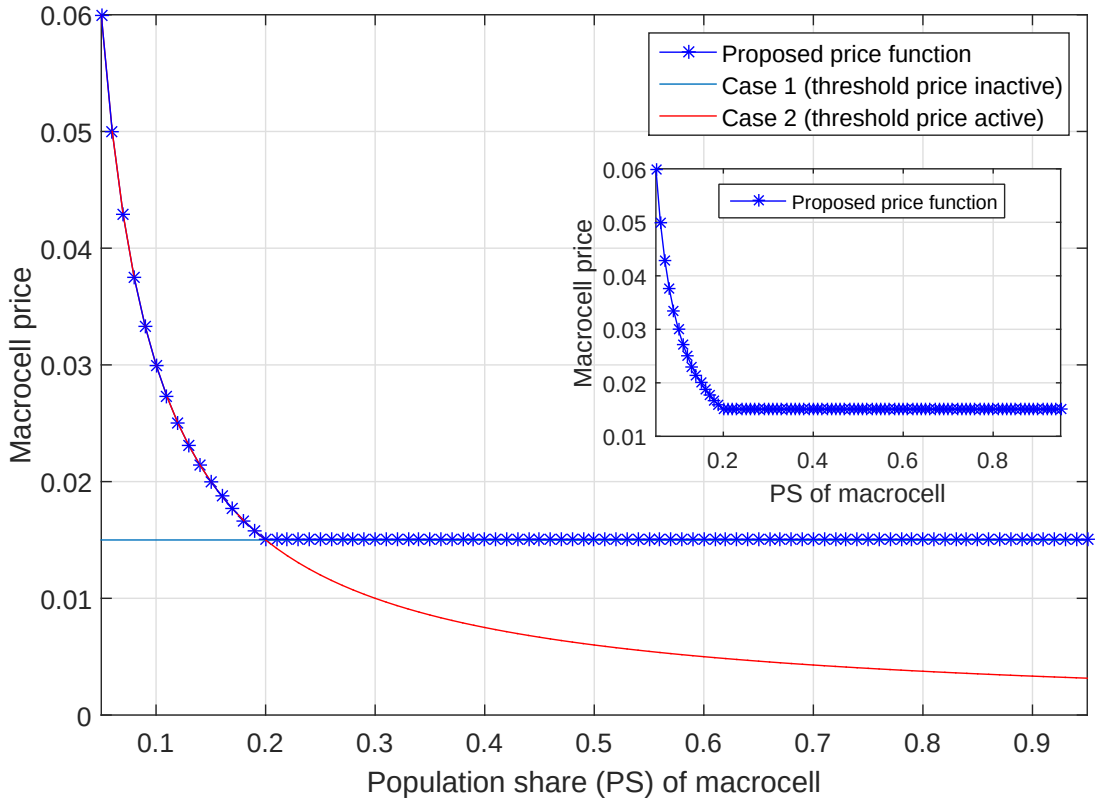


Figure 3.3: Illustration of the macrocell price function.

such as the rate-threshold R_τ and the small-cell interference price C_i shape the network dynamics by influencing the behavior of MUs. Before analysing the results, the values of the parameters used in simulations are listed: $\sigma = 0.1$, $\omega_M = 1$, $\omega_S = 0.9$, $R_S = 0.6$, $R_M = 1$, $C_f = 0.30$, $C_a = 0.002$, $C_i = 0.0005$ and $N = 100$.

3.5.1 Analysis of the Macrocell Cost Function

The characteristics of the proposed macrocell price function are obtained by simulating (3.1). The results demonstrate the mechanism of threshold price activation when the macrocell PS increases. The macrocell price function is plotted in Fig. 3.3 and compared against case 1 and case 2, using the value of $R_\tau = 0.05$. Case 1, plotted by the red line,

denotes the scenario where the price function increases or decreases linearly with respect to the average data rate achieved by the MUs. The macrocell adopts the linear pricing strategy as long as the macrocell PS remains below 0.2. Case 2, as shown by the blue line in Fig. 3.3, denotes the condition where the macrocell adopts the threshold price, the value of which can be scaled by using R_τ. As soon as the PS increases above 0.2 the proposed price function switches to case 2. The inset in Fig. 3.3 exhibits the proposed price function to show clearly the triggering point of the threshold price. The activation of the threshold price can be controlled by scaling R_τ for different PS settings. Later we validate by appropriate simulation that this threshold pricing strategy indeed influences MUs to connect with the small-cell tier.

3.5.2 Analysis of Cell Selection Dynamics

Payoff Dynamics

Fig. 3.4 gives the cell selection dynamics of the evolutionary game in terms of payoff functions. We have calculated the payoffs for an MU for selecting either a small-cell or the macrocell and plotted them against PS. Note that $R_\tau = 0.05$ and $C_i = 0.0005$. The payoff for an MU for macrocell selection decreases for both cases with the increase of macrocell PS. An MU's payoff for both case 1 and case 2 declines sharply as PS increases and intersects the small-cell payoff at a PS of 0.91 and 0.88, respectively. Note that these are macrocell PSs, i.e. $X_M = 0.91$ for case 1 and $X_M = 0.88$ for case 2. The average payoff increases linearly for both cases and also intersects the small-cell payoff function when the macrocell PS is 0.91 and 0.88, respectively. These are the equilibrium points of the network selection game. Thus we observe a contrasting phenomenon while analysing the equilibrium of the evolutionary game. The macrocell PS at the equilibrium never changes for the linear pricing strategy. On the contrary changing the rate-threshold in

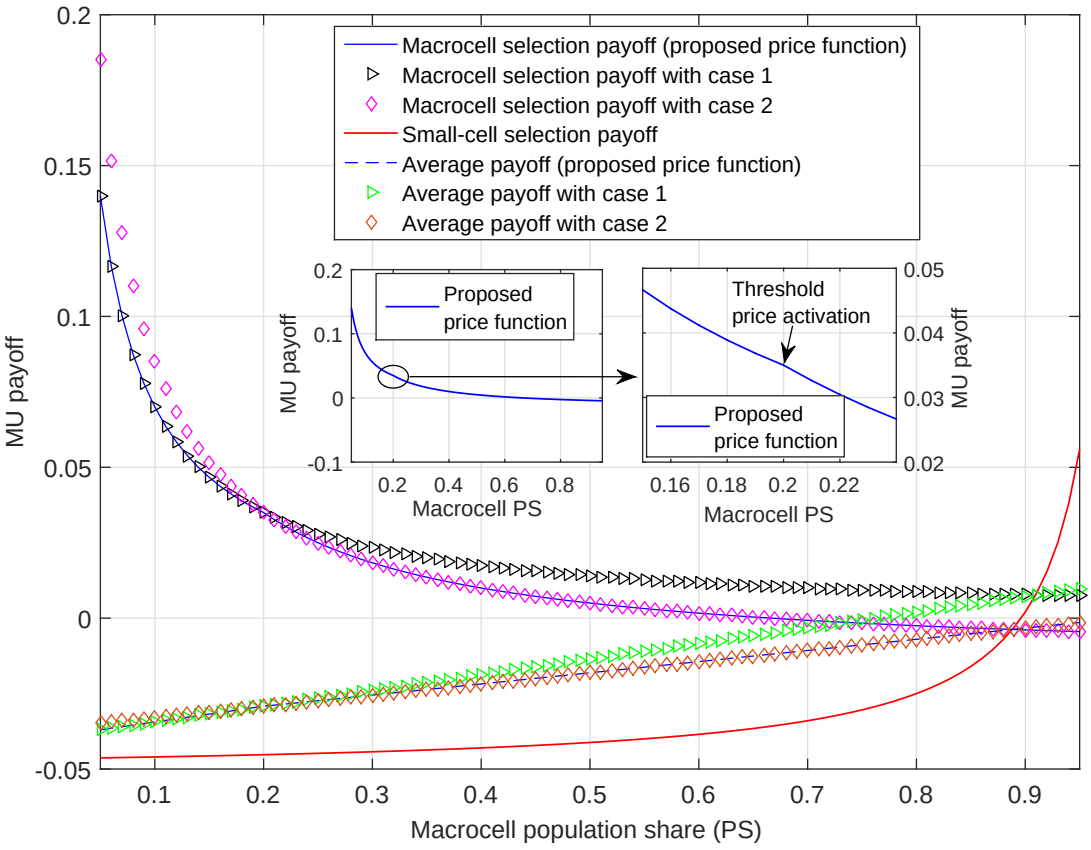


Figure 3.4: Analysis of the payoff functions.

the threshold pricing strategy changes the equilibrium to an another stationary point and allows the macrocell to offload a significant portion of the MUs to the small-cell tier. Later we show that by tuning R_τ a larger chunk of MUs can be offloaded to the small-cell tier. The insets in Fig. 3.4 show the sudden changes in MU payoff due to activation of the threshold price. Here, the activation of the threshold price offloads more than 3% of MUs to the small-cell tier.

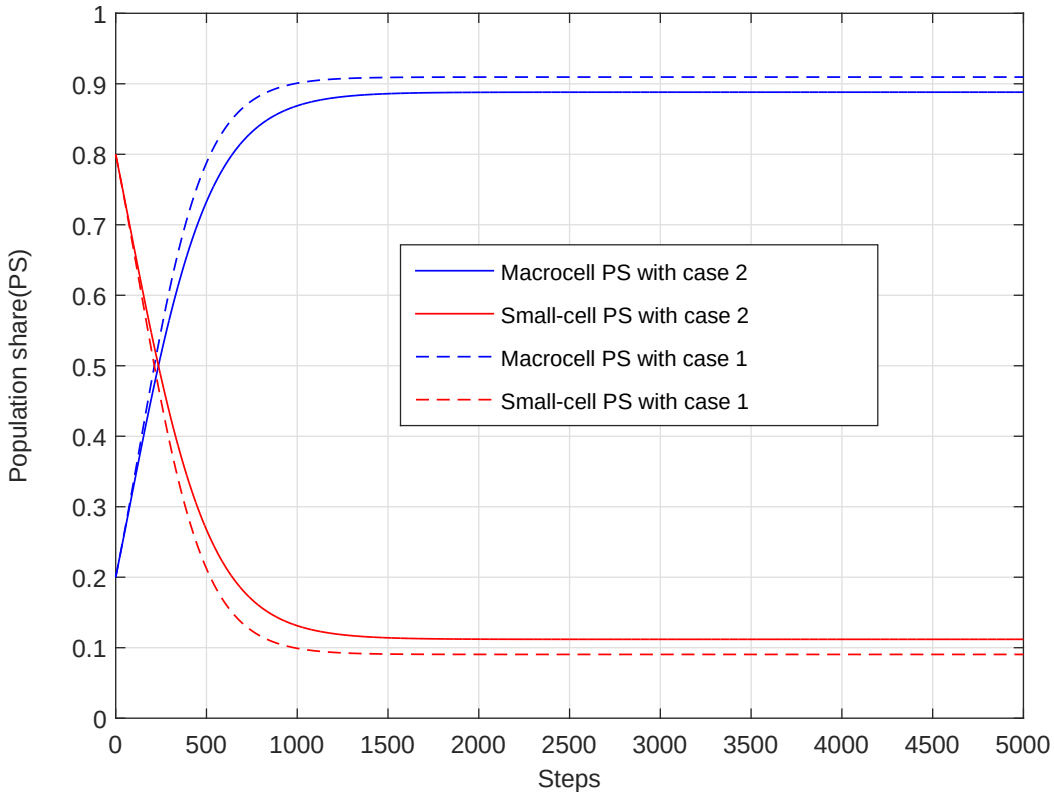
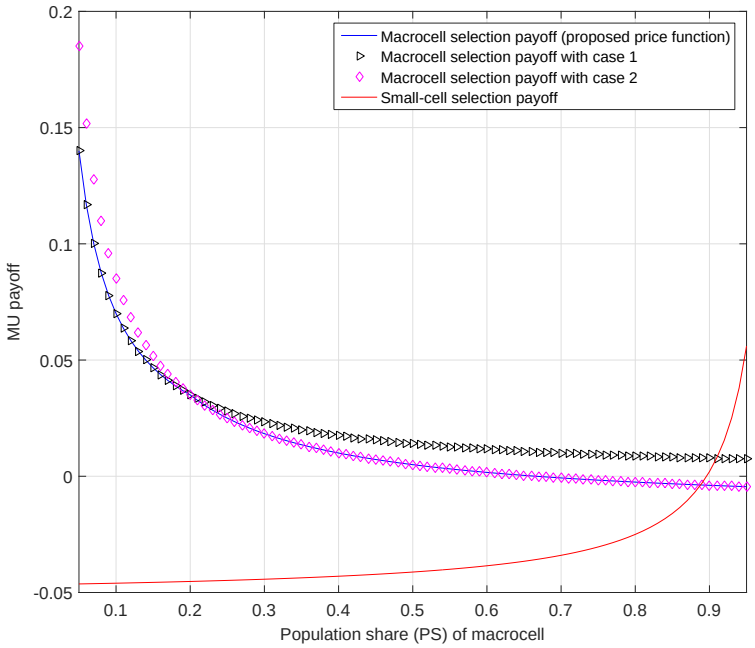


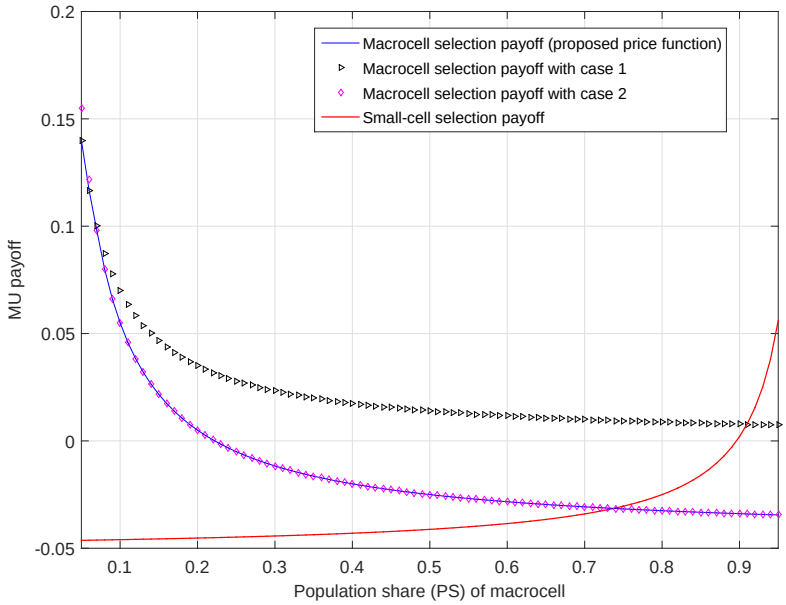
Figure 3.5: Cell selection dynamics by convergence.

Convergence

In this section, the convergence of the cell selection dynamics of the proposed evolutionary game is analysed for the corresponding cases using the parameter values $R_\tau = 0.05$ and $C_i = 0.0005$. In both cases, the replicator dynamics are simulated with the PS of both macrocell and small-cells plotted against time, given by Fig. 3.5. This gives us an idea how MUs reach a stationary solution and are distributed throughout the two-tier HetNet. At the start of the simulation the macrocell PS is 0.2. It is observed that the cell selection game approaches convergence as soon as its PS reaches 0.91 and 0.88 for case 1 and case 2, respectively. Note that the value of σ affects the rate of convergence. By increasing the value of σ a faster convergence can be expected.



(a) when the rate-threshold is 0.05,



(b) when the rate-threshold is 0.15,

Figure 3.6: Effect of the rate-threshold on MU payoff.

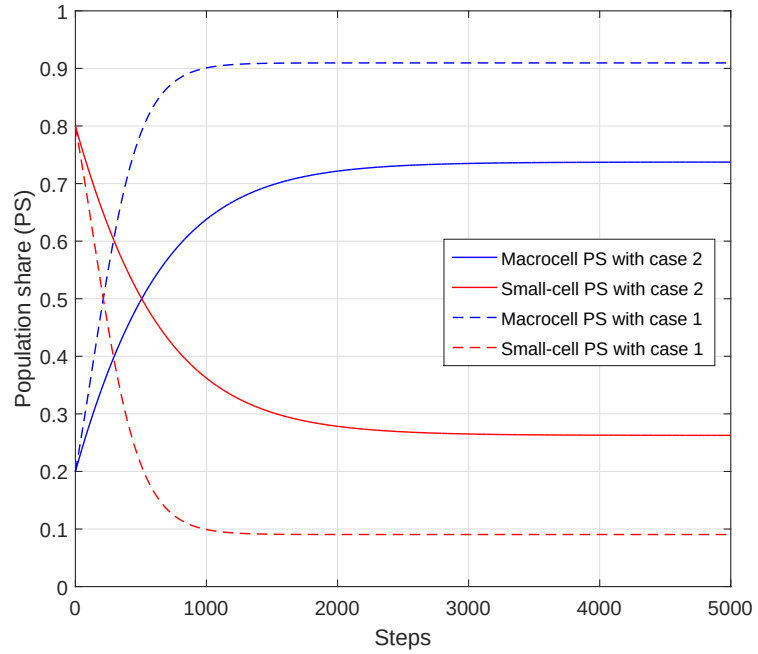
3.5.3 Effect of Rate-threshold

In this section, an extensive analysis is presented of the effect of the rate-threshold R_τ on the MU payoff and the network population share. The simulation results provide an insight on how R_τ influences an MU's decision in making network selection.

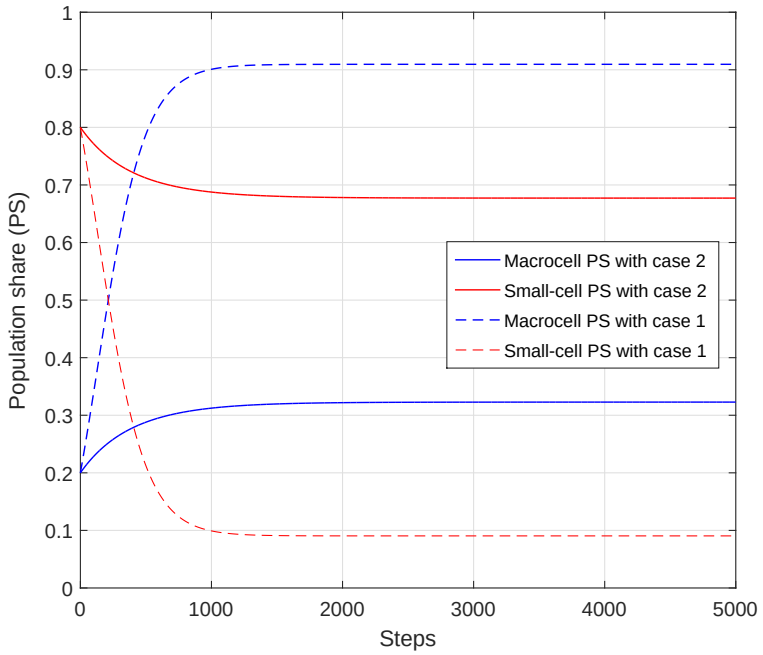
Payoff Functions

Fig. 3.6a and Fig. 3.6b show the reduction in the MU payoff for both cases. However, activation of the threshold price ($C_f R_\tau$) lowers the MU payoff further. The value of the threshold price is scaled by R_τ. As a result, MUs tend to associate with small-cells. Fig. 3.6a shows that activation of the threshold price lowers the macrocell PS by 3%. Fig. 3.6b shows a scenario where the macrocell payoff for case 2 is significantly lower than the payoff achieved with case 1. Note that $R_\tau = 0.15$ is used for simulation. The high value of the threshold price reduces the MU payoff so that a larger chunk of MUs select small-cells. Observe that the macrocell PS is 0.74 when the threshold price is active. Therefore, by increasing R_τ the macrocell is able to offload 15% more MUs to the small-cells compared to case 1 (linear pricing strategy), where the macrocell PS is 0.91.

The effect of the rate-threshold is now understandable by comparing the payoffs of case 1 and the proposed price function from Fig. 3.6a and Fig. 3.6b. Increasing the rate-threshold decreases the payoffs for both case 2 and the proposed price function. Also, by increasing the rate-threshold three times, from $R_\tau = 0.05$ to $R_\tau = 0.15$, more than 15% of the MUs can be offloaded to the small-cell tier.



(a) when the rate-threshold is 0.15,



(b) when the rate-threshold is 0.25,

Figure 3.7: Effect of the rate-threshold on PS.

Table 3.2: The effect of R_τ on network dynamics

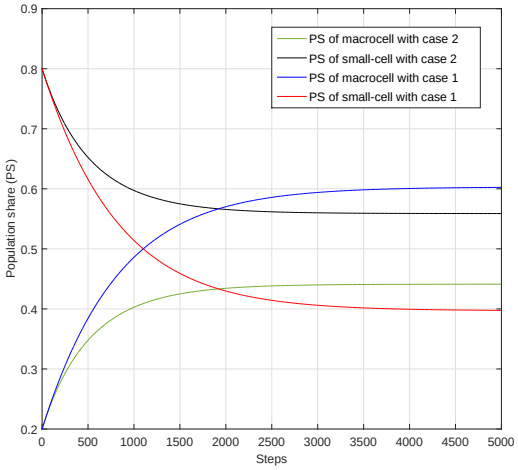
R_τ	Payoff (proposed cost function)	Population share (proposed cost function)
0.05	-0.0038	0.89
0.10	-0.0183	0.84
0.15	-0.0312	0.74
0.20	-0.0408	0.52
0.25	0.0441	0.32

3.5.4 Effect of Small-cell Interference Pricing on Network Selection Dynamics

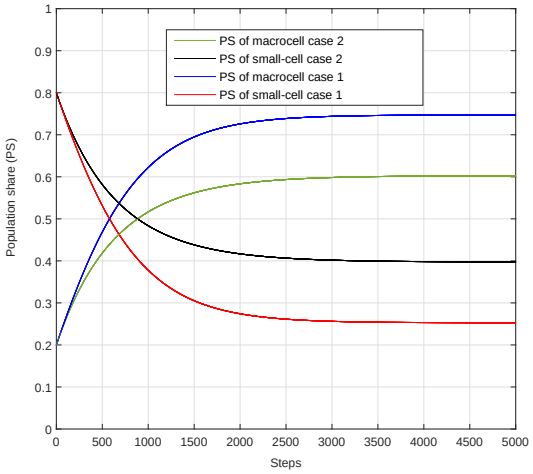
Analysis of Population Share

Now we analyse the effect of the rate-threshold on the macrocell's PS. Fig. 3.7 shows the convergence of the replicator dynamics for corresponding functions. The macrocell's PS, with the proposed price function for the rate-threshold of 0.15 and 0.25, are 0.74 and 0.32 respectively. On the other hand, case 1 (linear pricing strategy) can not affect the macrocell PS. We find that the macrocell PS with case 1 is 0.91 for all rate-threshold settings.

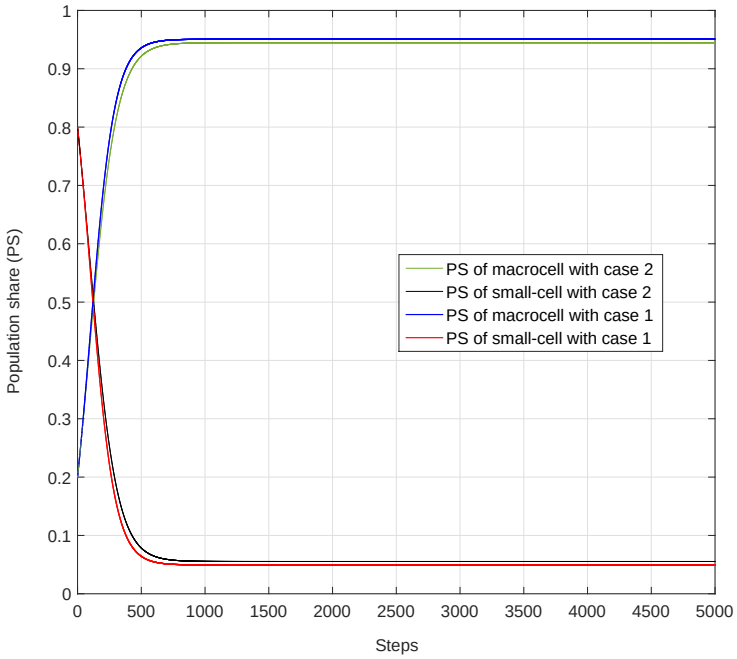
The payoffs and PS are listed in Table 3.2 for different values of R_τ to emphasise the effect of R_τ on the network dynamics as well as how the tuning of R_τ offloads MUs to the small-cell tier. Here the findings about small-cell interference pricing and how it shapes the network dynamics are presented. The goal of implementing interference pricing as a small-cell pricing strategy is to restrict distant MUs from connecting with a small-cell. Small-cells are only allowed to increase their transmission power to a fixed level so that nearby MUs can connect with them.



(a) $C_i = 0$,



(b) $C_i = 0.0001$,



(c) $C_i = 0.001$,

Figure 3.8: Effect of the small-cell interference price.

In the studies conducted R_τ was fixed at 0.05 and C_i was varied. As shown in Fig. 3.8a, the macrocell PS for case 1 and case 2 are 0.6023 and 0.4412 respectively at $C_i = 0$. It is

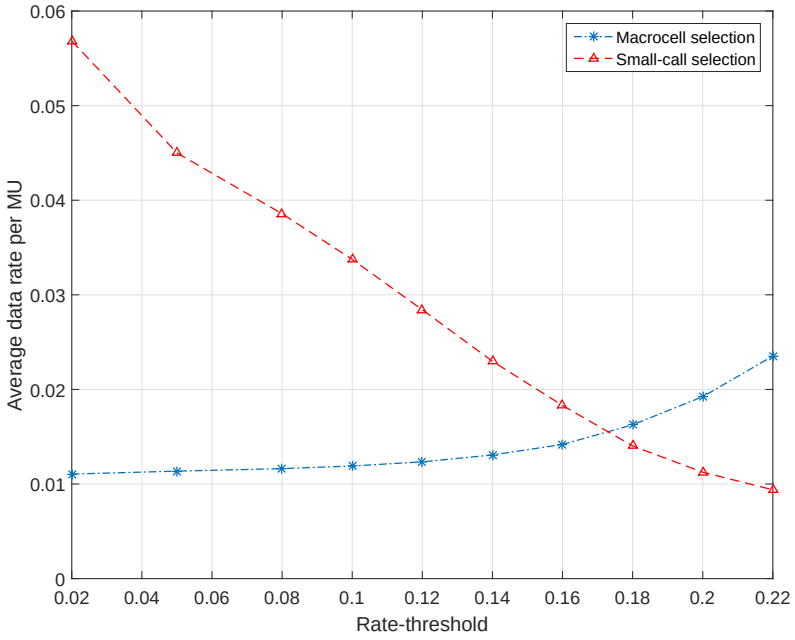


Figure 3.9: Effect of the rate-threshold on the average data rate per user.

observed that a large group of users connect to small-cells if the interference price is not activated. For $C_i = 0.0001$, the PSs of the macrocell for case 1 and case 2 respectively are 0.747 and 0.6022. As the interference price increases, a change in the network dynamics is seen with more users tending to join the macrocell. Note that the difference in the macrocell PS between case 1 and case 2 is due to the threshold price, as shown in Fig. 3.8a, Fig. 3.8b and Fig. 3.8c. Due to the threshold price, the macrocell PS for case 2 is significantly lower than in case 1. As C_i is increased tenfold from, 0.0001 to 0.001, the macrocell PS for case 1 and case 2 becomes 0.9506 and 0.9447, respectively. The jump in the macrocell PS for both cases in Fig. 3.8c is due to the high interference price charged by a small-cell. An interesting observation is that, as the small-cell interference price becomes higher, the influence of the threshold price on the macrocell PS reduces. As a result the macrocell PS achieved for case 1 and case 2 is almost the same.

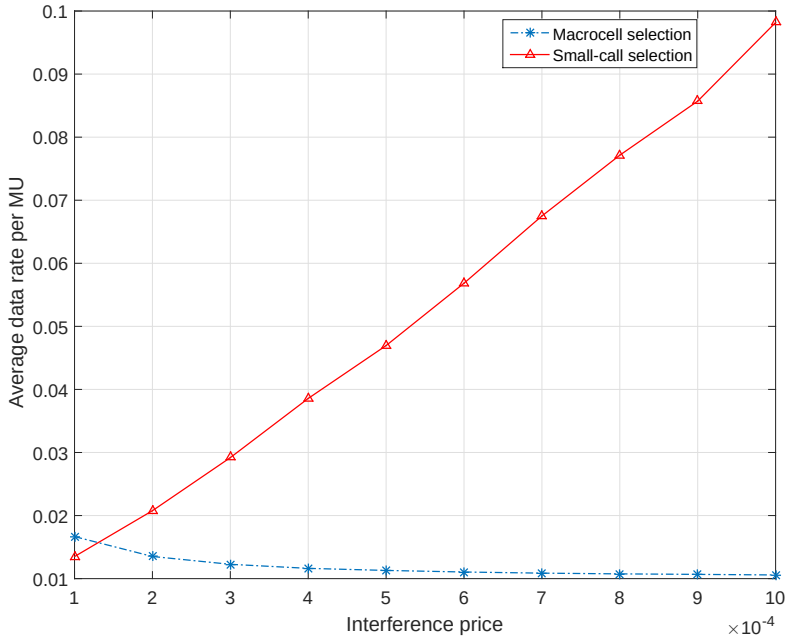


Figure 3.10: Effect of the interference price on the average data rate per user.

3.5.5 Average Data Rate Performance

Here we investigate the effect of the rate-threshold and the interference pricing on the average data rate performance of the MUs. Fig. 3.9 shows the impact of the rate-threshold on the average data rate per MU. We run the simulation by varying the rate-threshold only and record each of the equilibrium points, i.e. the macrocell PS corresponding to that rate-threshold. We then compute the average data rate per MU with respect to the previously recorded macrocell PS and plot them against the rate-threshold. The majority of MUs choose the macrocell at the lower value of rate-threshold and it is the lowest at $R_\tau = 0.02$. As a result the average data rate per user for macrocell association is very low. On the contrary the average data rate per user for small-cell association is the largest for $R_\tau = 0.02$. We observe that a higher value of rate-threshold pushes more MUs to small-cells. As a result the average data rate per MU drops considerably for small-cell association.

Fig. 3.10 shows the effect of interference pricing on the average data rate per MU. In this case we vary the interference price and record the shift in the equilibrium. We observe that the average data rate per MU for small-cell association remains lowest when the value of the interference price is low, i.e. when $C_i = 0.0001$. This is due to the fact that a low interference price motivates more users to select small-cells for association. However, a higher value of interference price will lead to a higher number of users to be associated with macrocells. Thus MUs who select the macrocell will have a lower average data rate. On the contrary, the average data rate of MUs for small-cell association will increase with higher interference price.

3.5.6 Performance Benchmark

The following performance measure serves as a benchmark in our analysis. In order to measure the effectiveness of our proposed price-based offloading scheme we introduce a new parameter called the *offloading efficiency*. The offloading efficiency is defined as the ratio of the offloaded MUs compared to the total MUs, i.e.,

$$\text{offloading efficiency in}\% = \frac{\text{total MU} - \text{equilibrium macrocell population}}{\text{total MU}} \times 100. \quad (3.53)$$

The offloading efficiency indicates the performance of both pricing schemes in achieving the desired offloading performance in terms of optimum population share in both tier. Fig. 3.11 shows the effect of the rate-threshold in offloading users to the small-cell tier and the effect of the interference price in accommodating MUs to small-cells. We find that the offloading efficiency of the macrocell depends on the rate-threshold of the macrocell pricing strategy. We calculate the offloading efficiency by collecting the equilibrium population data for each rate-threshold setting by keeping the interference price constant. As the rate-threshold increases, the offloading efficiency also increases, i.e. more MUs select small-cell tier for association. We see that, at rate-threshold 0.22, more than 50% users select small-

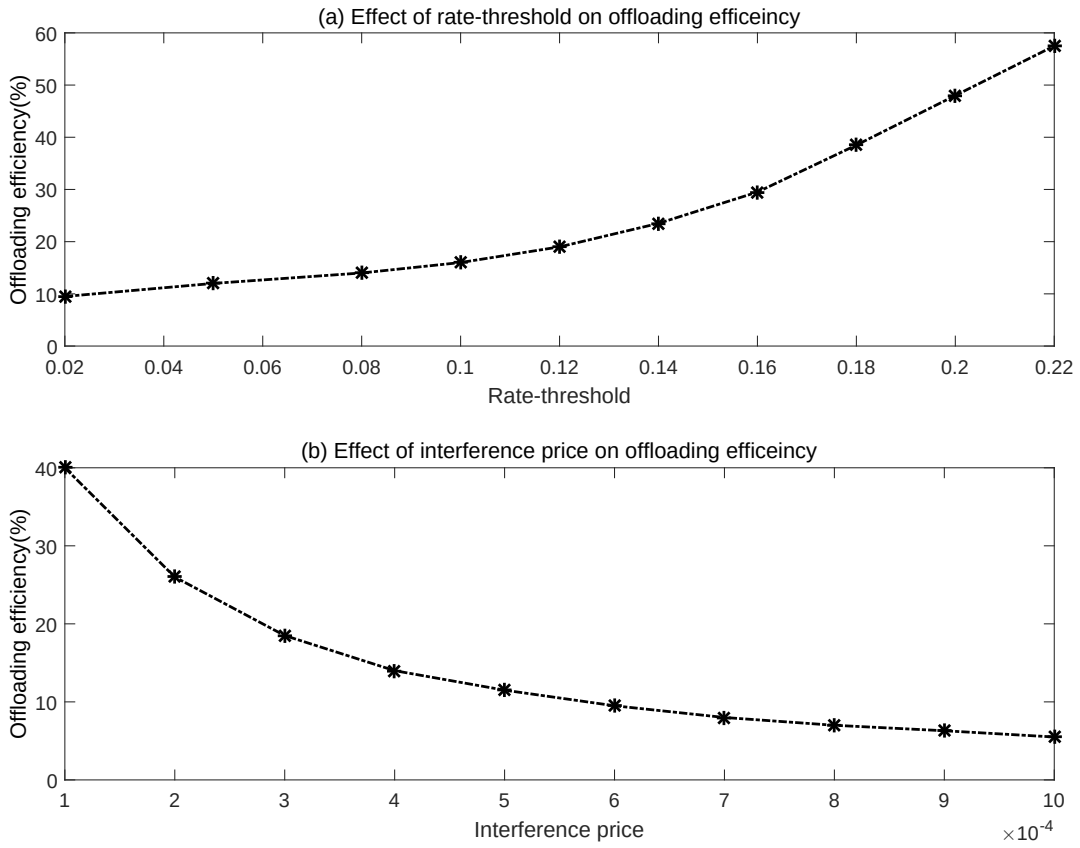


Figure 3.11: Effect of the rate-threshold and interference price on the offloading efficiency.

cells. Thus the potency of our proposed macrocell pricing strategy can be understood in terms of the rate-threshold. The rate-threshold not only triggers the threshold price where each MU pays more than the linear price, but also provides the macrocell a much needed control mechanism to offload a portion of its population to small-cell tier. On the contrary, increasing the interference price has a negative effect on a user's small-cell selection strategy. At a higher interference price more users select the macrocell. We calculate the offloading efficiency by keeping the rate-threshold constant while we collect the equilibrium population data for different interference price-settings. Thus there exists a trade-off between the rate-threshold of the macrocell pricing strategy and

the interference price of the small-cell pricing strategy. This trade-off can be utilised to have an optimum PS in the macrocell and small-cell tiers.

3.6 Summary

In this chapter, a new macrocell price function is proposed, which combines a linear pricing strategy with a threshold pricing strategy. The pricing strategy is devised in such a way that a macrocell is able to offload a proportion of its users to small-cells effectively. The small-cell pricing strategy is formulated in a way that it restricts a large group of MUs from joining the small-cell tier. The introduction of a rate-threshold in the threshold pricing strategy provides a macrocell with a degree of freedom for scaling the number of MUs in a two-tier HetNet. Furthermore the rate-threshold and the interference price of a small-cell can be optimised to have the optimum PS in each tier of the HetNet.

Chapter 4

A Nash Bargaining Game for Cooperative Content Association in 5G HetNet

4.1 Chapter Introduction

Content caching via small-cells is a special case of user offloading where users download large data files, for example, popular video contents, gaming contents etc., to reduce the burden on macrocells originating from redundant downloads. Realising the potential of SBS-based content caching in 5G networks, this chapter discusses a novel caching strategy which is developed utilising the locally available caching opportunities provided by small-cells. Existing caching strategies mainly rely on small-cells for hosting and caching smaller contents. On the other hand, for retrieving larger contents, the requested contents are either downloaded from cloud storage or the requests are redirected to a nearby macrocell. Both cases result in a higher delay, which works as a motivating factor to propose a caching strategy that increases small-cell utilisation for caching purposes. Therefore a

novel caching policy is proposed that aims to reduce the caching load on MBSs as well as the burden on their backhaul, due to the redundant transmission of popular contents.

A content-caching strategy involves two decision-making criteria: i) content placement and ii) association for content retrieval. Therefore a content-caching policy can be devised either from an SBS perspective, i.e. using the serving nodes to deal with the content placement, or from a user perspective, i.e. using the nodes for service, based on a suitable base station selection, which provides the user with the lowest delay. Thus a content-caching problem can be formulated in terms of a classic generalised assignment problem (GAP) [111], where placing a content is analogous to a GAP, or in terms of a classic facility location problem (FLP) [112], where association with a cache is analogous to an FLP. Both GAP and FLP are jointly considered and solved in [113]. Traditionally, the classes of GAP and FLP are combinatorial optimisation problems and therefore hard to solve.

The proposed content-caching model deals with cache association for users. A novel delay-based utility function is adopted to incorporate a delay restriction on the user's content-retrieval delay. To associate with an SBS, a user's content-retrieval delay must remain under a certain threshold. Choosing an appropriate delay threshold is very important in the context of maximising SBS association. Though allowing the delay threshold to be content specific seems a practical choice, it will significantly increase the control channel overhead due to the additional information requirement for computing a user-specific threshold delay. As a result, we assume that the delay threshold will be the average delay requirement of a user when a content is downloaded from a macrocell. This assumption allows the content-caching model to remain tractable even when the number of users increases in the HetNet and it also provides a holistic view about maximising SBS-based content association. The introduction of the threshold delay in the caching model also provides an insight about a delay upper bound for 5G HetNet. In addition

the introduction of NBG ensures that users cannot exhibit selfish behaviour. Thus a successful negotiation results in a fair bandwidth allocation which in turn translates into a delay reduction for users. For example, a one-millisecond end-to-end delay is envisaged for 5G networks, which can be used as a threshold delay parameter in our caching model to compute association of users with SBSs. The premier features of the proposed caching model are the maximisation of SBS utilisation for content caching and its ability to numerically analyse how many users can associate with an SBS under different delay restrictions, while most of the existing caching models aim to reduce the content download delay [73], [72] for a user. In Section 4.5.4 two special cases have been analysed by introducing a stringent and a flexible delay concept. The stringent delay concept (suitable for real-time application) is introduced for services where the required delay threshold has to be satisfied to avoid performance variation. On the other hand, a flexible delay concept (suitable for video file caching) is introduced and analysed. How these delay thresholds can affect user association has also been analysed. The interaction among users and an SBS for cache-association is modelled using the well-studied game-theoretic concept called the NBG. In the cache-association game, rational users will collaborate if the collaboration incurs less delay for them than working individually. In an NBG-based caching model, a user therefore wants to attain a greater utility via collaboration with other users, which is then translated into a fair and unique optimal solution.

A cache-association decision is modelled using a binary decision variable. The content-retrieval delay is a function of the allocated bandwidth, so that the content-retrieval delay also depends on the bandwidth. Therefore the content-caching system is formulated into an NBG model that jointly maximises the bandwidth and cache-association decisions. By carefully analysing, we find that the formulated NBG is a combinatorial integer programming problem. Firstly the constraint relaxation approach is applied to solve the NBG, which provides a centralised solution of bandwidth and cache-association decisions.

Table 4.1: List of Important Symbol Notations

Symbol	Description
$\mathcal{I}$	Set of SBSs
$\mathcal{M}$	Set of users
$\boldsymbol{t}$	Vector of all delay variables
$\boldsymbol{x}$	Vector of all cache-association variables
$\mathcal{Q}$	Set of all contents cached by SBS
C_q	Size of the q-th content
$\boldsymbol{w}$	Vector of bandwidth allocations
w_{mi}	Allocated bandwidth to m-th user by i-th SBS
Υ_i	SNR of the i-th SBS
t_τ	Delay threshold
t_{mi}	Delay for m-th user
W_i^{max}	Maximum bandwidth of SBS i
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\mu}$	Set of non-negative Lagrange multipliers

Secondly, the Lagrange relaxation approach is applied, which makes it possible to decompose the problem into individual user problems. Finally, the subgradient method is used to iteratively compute cache association and bandwidth allocation for each user in a distributed way.

4.2 Network Model and Assumptions

4.2.1 Notation Summary

The notation used is summarised in Table 4.1.

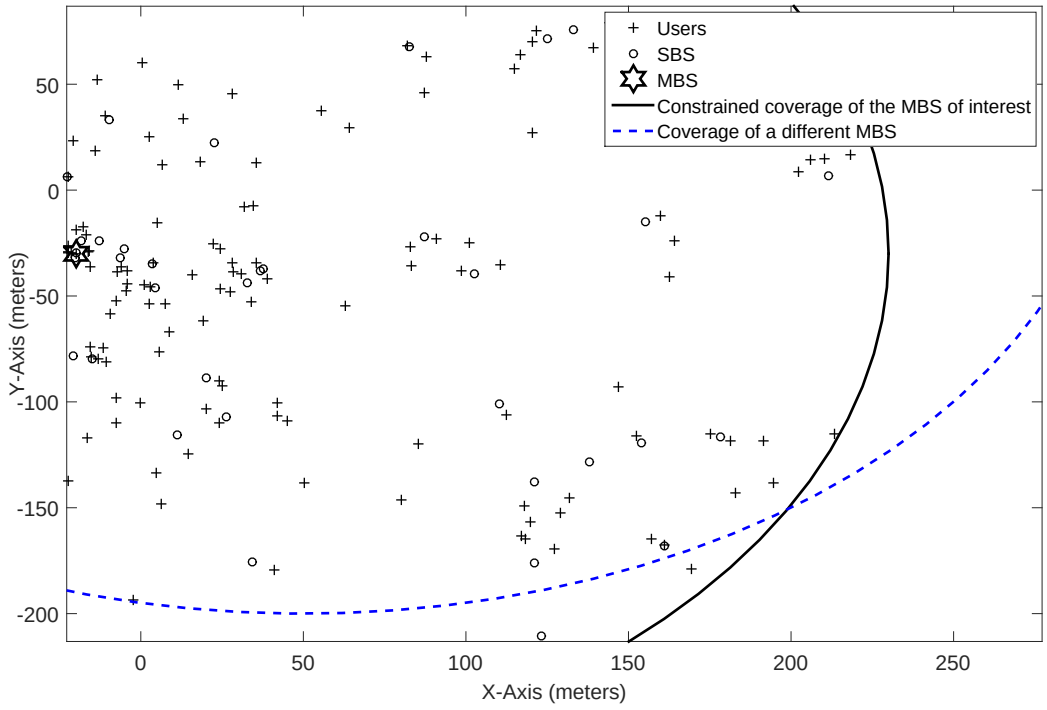


Figure 4.1: Network schematic showing the constrained area of the HetNet.

4.2.2 Elements of Wireless Networks

Fig. 4.1 shows a portion of the HetNet consisting of a single MNO, where the MNO deploys one MBS, and a large number of SBSs. By fixing the coordinates of the MBS, a total of 50 SBSs are placed randomly inside the coverage area of the MBS. Users are also randomly placed inside the macrocell coverage to capture a realistic scenario, where some of them have the opportunity to associate with SBSs to download contents. To have a better understanding, a tiered content-caching architecture is visualised in Fig. 4.2, where a user first tries to connect with an SBS to retrieve a content. It tries to associate with the nearest SBS via negotiation with other competitors (users) with a view to reducing its own content-download time. If a negotiation is successful, a user will associate with an SBS. Otherwise it will try to retrieve the content from another MBS. Thus the cooperation via

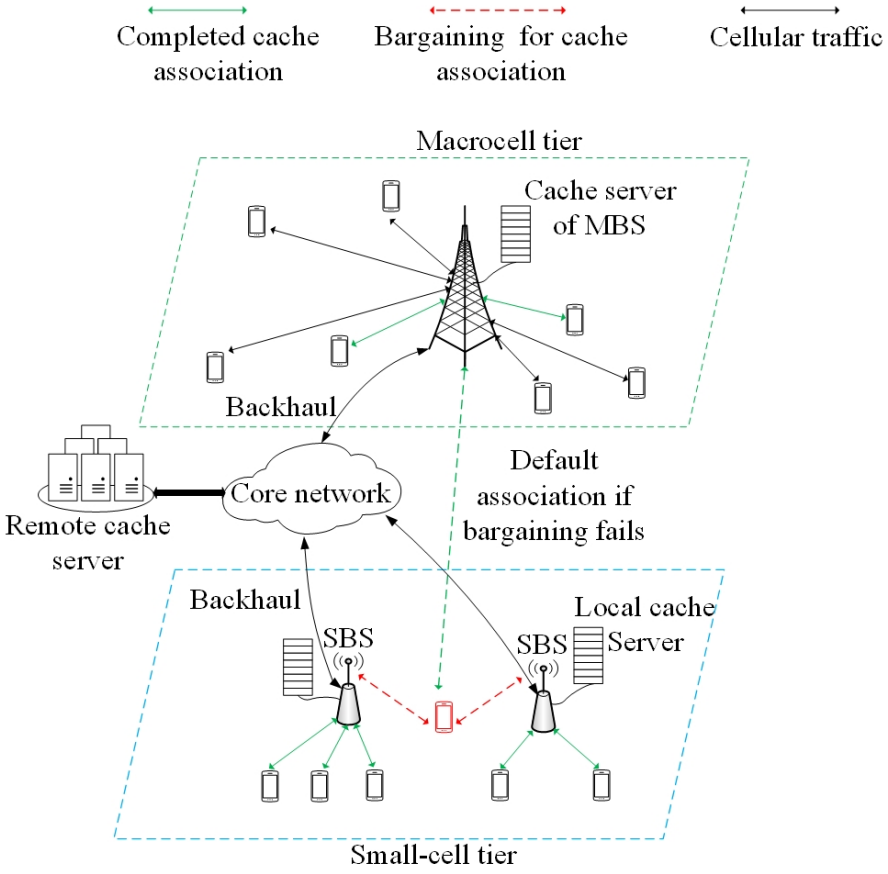


Figure 4.2: Illustration of a content-caching system in a two-tier HetNet.

negotiation provides a win-win scenario for both MNO and users, where an MNO offloads a significant portion of the MBS's content-caching load to SBSs. On the other hand the delay for each user is reduced since they are being served locally. We assume that the MBS and SBSs adopt LTE in their physical layer to provide services to users. The set of all SBSs is denoted by $\mathcal{I} = \{1, \cdots, I\}$, where $i \in \mathcal{I}$ denotes the i-th SBS. We denote the set of users as $\mathcal{M} = \{1, \cdots, M\}$, where $m \in \mathcal{M}$ denotes the m-th user. We assume that the MNO deploys SBSs very close to each other so that they can serve as many content requests as possible. The MBS and SBSs are connected with the core network, with a high capacity, and a limited-capacity wired backhaul respectively.

Since spectrum is a scarce resource, spectrum sharing is preferred for SBS operation.

Therefore it is assumed that the MBS and SBSs share the same spectrum (sub-bands), i.e. they are deployed in an underlay-type spectrum-access manner. In addition, an interference coordination mechanism is necessary among the MBS and SBSs to avoid inter-tier and intra-tier interference in the downlink channel [114], [56], [115]. Under this assumption, we can use the signal-to-noise-ratio (SNR) instead of the signal-to-interference-plus-noise-ratio (SINR) [72]. In [72], each SBS or each device transmits with the same SNR, however the received SNR of a user depends on the distance between the SBS and that user. We adopt a similar SNR model in this chapter for analytical convenience. We assume that all SBSs transmit at the same power level [116] and they have the same coverage [117]. In addition, no random shadowing and identical receiver noise are assumed so that the same cell-edge SNR can be ensured [117] in all SBSs. This assumption can be removed by devising a joint caching and transmission policy where an SBS individually tunes its transmission property, such as transmission power, to keep the SNR above a minimum threshold, which in turn reduces the download delay for each user. We denote the SNR of the i-th SBS as Υ_i.

Given that the SNR is fixed, the data rate of a user depends on the amount of allocated bandwidth. On the other hand the amount of bandwidth which is going to be allocated depends on the number of users arriving at an SBS in a single time-slot. To elaborate on this phenomenon, let us consider an example of allocating bandwidth in an LTE wireless network. The allocation of bandwidth in a practical LTE wireless network depends on the number of assigned physical-resource blocks (PRBs)[1] per user. For example, consider an SBS-based caching system, equipped with a commercial LTE network with a system bandwidth of 10 MHz, which has a maximum of 50 PRBs [119]. If there are 10 requests at an SBS from 10 users, each user can have a maximum allocation of 5 PRBs. On the

[1]A PRB is the smallest unit of resource in an LTE physical layer, with a dimension of 1 millisecond in the time domain and 180 kHz in the frequency domain [118]

other hand, each user can have a maximum allocation of 2 PRBs when 25 users place content-download requests. By adopting priority-based PRB allocation, high-priority users can be assigned more PRBs than low-priority users. Note that both single-PRB per-user assignment [119] as well as multi-PRB per-user assignment [120], [121] have been discussed in the literature. Thus the content-retrieval delay depends on the number of associations and the allocated bandwidth for each user.

4.3 Problem Formulation

In this section, we will investigate a problem for cache association with an assumption that SBSs independently place a fixed amount of contents. We also assume that all contents have the same size [72], [73] to maintain the tractability of the proposed NBG-based caching model. Therefore we will occupy ourselves analysing a cache-association problem from a user-association and a bandwidth-allocation perspective. However, Chapter 6 will investigate the cache-association problem from a joint cache-placement and cache-association perspective.

Consider that all contents placed by the SBSs are the same. This is possible when there exists a global library. The global content library is defined by the set $\mathcal{Q} = \{1, \cdots, Q\}$. $|\mathcal{Q}|$ is the cardinality of set $\mathcal{Q}$ that defines the total contents that can be independently cached by an SBS. Let C_q denote the size of the q-th content in bytes, $q \in \mathcal{Q}$. Suppose that the m-th user wants to retrieve the q-th content from SBS i. To calculate the maximum number of associations, we first define a cache-association vector as follows:

$$\boldsymbol{x} = (x_{mi} : m \in \mathcal{M}, i \in \mathcal{I}), \tag{4.1}$$

where $x_{mi} \in \{0, 1\}$ is a binary decision variable which defines whether the m-th user requesting any content q is going to associate with the i-th SBS or not. Assume that W_i^{max} is the maximum bandwidth of SBS i. Suppose that m-th user gets a bandwidth

w_{mi} after bargaining. Therefore the data rate of user m for content-download purposes is given by:

$$R_{mi} = w_{mi} \log(1 + \Upsilon_i). \tag{4.2}$$

Delay is an important metric in the context of a content-caching system and defines how fast a user can retrieve a content after association with a base station. Using (4.2) the delay for user m can be computed. The delay for retrieving the q-th content is:

$$t_{mi} = \frac{x_{mi} C_q}{w_{mi} \log(1 + \Upsilon_i)}, \tag{4.3}$$

Notice that the delay is dependent on the content size C_q, the number of associations and the allocated bandwidth w_{mi} for the m-th user. In order to associate with SBS i, the delay requirement for the m-th user must satisfy:

$$(t_\tau - t_{mi}) > 0, \tag{4.4}$$

where t_τ is the threshold delay, which is computed in terms of the average download delay per user when associating with an MBS, and t_{mi} is the delay for the m-th user for retrieving the q-th content from the i-th SBS. Notice that we only consider the transmission delay for a content-caching model, i.e. the content download delay is the main performance metric for determining a cache association. Though (4.4) intuitively avoid congestion by assuming a strict threshold delay, queuing delay and other kinds of delay is avoided in the delay formulation to maintain solution tractability. However it is our aim to include these delays in the future work of this chapter. Therefore the m-th user will only associate with the i-th SBS when the content-retrieval delay satisfies (4.4), i.e. if the content-retrieval delay for SBS association is lower than the average download delay while the same content is being retrieved from an MBS. Now the delay vector can be defined as follows:

$$\boldsymbol{t} = (t_{mi} \colon m \in \mathcal{M}, i \in \mathcal{I}). \tag{4.5}$$

4.4 Cache-Selection Game: The Nash Bargaining approach

In this section, the cache-selection procedure is modelled into an NBG. In the NBG model, users are considered as players and SBSs' caches are considered as commodities. A question may arise why we are choosing cooperative NBG solutions over noncooperative game-theoretic solutions. Yang et al. [60] masterfully demonstrate the inefficiency of the Nash equilibrium (please refer to Fig. 1 of [60]) and the effectiveness of bargaining game-theoretic solutions. Furthermore the Pareto optimality and fairness axioms make the NBG a superior choice over non-cooperative solutions to formulate and solve our proposed SBS-based caching model. For more details on this issue, please refer to Chapter 2 of this thesis. The first step towards formulating an NBG-based optimisation model is to define an agreement utility function for the m-th user for associating with the i-th SBS. However a user may not be able to reach an agreement with other users (bargainers) for associating with the i-th SBS when the bargaining outcome is less than its minimum requirement. Therefore a user needs to consider the SBSs in its coverage for association so that it will increase its chance to associate with an SBS after failing to reach an agreement with an SBS. Thus bargaining could occur sequentially or concurrently, as explained in [92].

Fig. 4.3 serves as a motivating example of our proposed bargaining-based content-caching system. It captures the caching problem from the users' perspective, where each user selects an SBS that provides the lowest download delay, as well as from the SBSs' perspective where each SBS judiciously decides on placing contents considering the content popularity. We mainly discuss the cache-association problem in this chapter while the content placement issue is discussed in Chapter 6. Since the delay depends on the allocated bandwidth of each user and the bandwidth is shared among users, the content-retrieval delay of users is affected by the delay of other users in the same SBS. Fig.

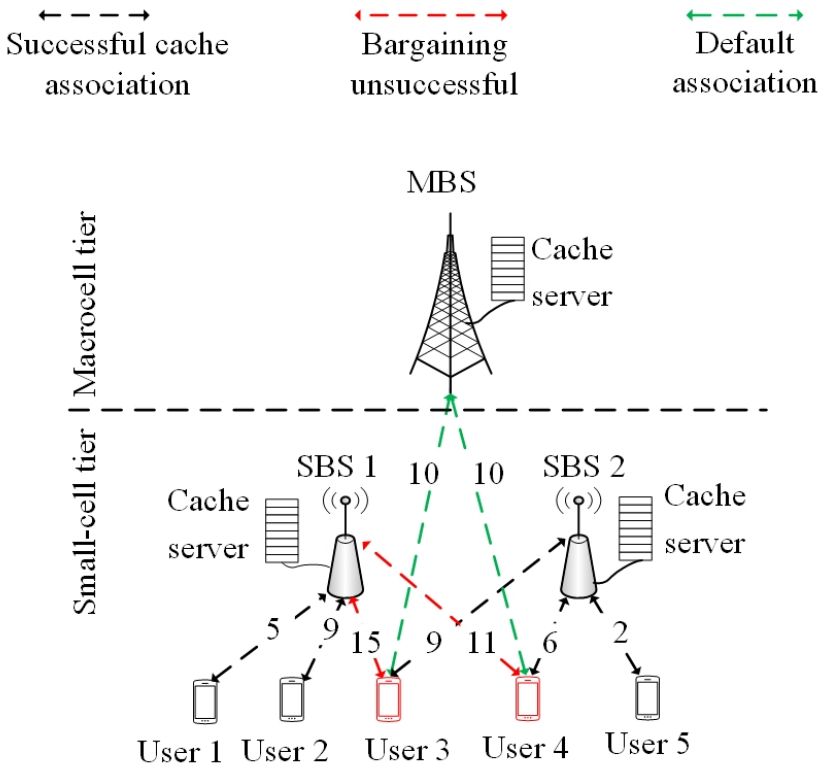


Figure 4.3: Illustrative example of a bargaining-based SBS association for content retrieval in a two-tier HetNet.

4.3 demonstrates a user-association procedure which adopts the concept of a bargaining-based association. It shows that an MNO deploys two SBSs with local caches. Each SBS is capable of serving a finite number of users depending on the size of their cache requests and their transmission bandwidth. For example, there are 5 users in the HetNet as shown in Fig. 4.3 and each user is trying to access either of the SBSs for downloading contents. The numerical value in each line demonstrates the delay which each user incurs while downloading their preferred content from the respective SBSs. Fig. 4.3 shows that the delay requirements of both user 1 and user 2, are satisfied. However both user 3 and user 4 face a higher delay to retrieve their contents, and bargain with other users for associating with either SBS 1 or SBS 2. The green dotted lines represent the expected average content-retrieval delay for both user 3 and user 4, when they download contents

from an MBS. This average delay is considered to be the threshold delay or default delay requirement for all users, which is required to be satisfied for association with a particular SBS. Due to the bandwidth restriction and/or the volume of requested contents, there is a 15 second and a 12 second delay for user 3 and user 4 respectively, if they associate with SBS 1. On the other hand, associating with SBS 2 yields those users a delay of 9 seconds and 7 seconds respectively. Certainly both users will select SBS 2, since the default delay requirement for SBS association is satisfied. Therefore we need to develop a cache-association algorithm that maximises users' chances to retrieve contents from the SBSs. Thus, in the NBG, users bargain with other users to associate with an SBS with a view to reducing the individual content-retrieval delays.

We adopt a delay-based utility function for the NBG formulation. Let $\mathcal{N}_m$ be the set of locally available SBSs around user m, where $\mathcal{N}_m \subset \mathcal{I}$. Here the assumption is that user m is under the coverage of a set of SBSs. Therefore user m can bargain for association with any of the SBSs $i \in \mathcal{N}_m$ but it can only associate with a single SBS. We define the agreement utility function for the m-th user as $U_m(t_{mi}) := (t_\tau - \sum_{i \in \mathcal{N}_m} t_{mi})$ for associating with the i-th SBS. The disagreement utility for the m-th user is $U_m^0 := t_{min} = 0$. The interpretation of the disagreement utility is that the m-th user will not associate with the i-th SBS if the delay condition is not satisfied, i.e. if the download delay is higher than that of an MBS. We can write the optimisation problem in the standard form of a

GNBS [86], [87]:

$$\underset{\boldsymbol{x},\boldsymbol{w},\boldsymbol{t}}{max} \quad \sum_{m\in\mathcal{M}} \log\left(U_m(t_{mi}) - U_m^0\right) \tag{4.6a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} x_{mi} = 1, m \in \mathcal{M}, \tag{4.6b}$$

$$\sum_{i\in\mathcal{N}_m} t_{mi} \leq t_\tau, m \in \mathcal{M}, \tag{4.6c}$$

$$w_{mi} \geq 0, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{4.6d}$$

$$\sum_{m\in\mathcal{M}} w_{mi} = W_i^{max}, i \in \mathcal{N}_m, \tag{4.6e}$$

$$x_{mi} \in \{0,1\}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{4.6f}$$

From (4.3) we can express w_{mi} as:

$$w_{mi} = (A_{mi} \cdot \frac{x_{mi}}{t_{mi}}), \tag{4.7}$$

where $A_{mi} = \frac{C_q}{\log(1+\Upsilon_i)}$ is a constant. Substituting the values of the utility functions and using (4.7) we obtain the following optimisation problem:

$$\underset{\boldsymbol{x},\boldsymbol{w},\boldsymbol{t}}{max} \quad \sum_{m\in\mathcal{M}} \log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) \tag{4.8a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} x_{mi} = 1, m \in \mathcal{M}, \tag{4.8b}$$

$$\sum_{i\in\mathcal{N}_m} t_{mi} \leq t_\tau, m \in \mathcal{M}, \tag{4.8c}$$

$$w_{mi} = \left(A_{mi} \cdot \frac{x_{mi}}{t_{mi}}\right), i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{4.8d}$$

$$\sum_{m\in\mathcal{M}} w_{mi} = W_i^{max}, i \in \mathcal{N}_m, \tag{4.8e}$$

$$x_{mi} \in \{0,1\}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{4.8f}$$

The optimisation problem presented by (4.8a)-(4.8f) is our generalised Nash bargaining model, where each user's utility is generated by successfully downloading the desired

content. Each user joining the game, i.e. willing to associate and retrieve a content, therefore cooperates with a view to maximising their utility. If the utility preference has not been met, the user can try to retrieve that content from elsewhere. Constraint (4.8b) means that a user can only associate with one SBS. Constraint (4.8c) explains that the delay must not exceed the threshold delay. Constraint (4.8d) explains that the bandwidth requirement has to be satisfied. On the other hand, constraint (4.8e) imposes a bandwidth constraint for all users. It states that the total bandwidth allocation of an SBS is bounded by its maximum bandwidth. (4.8f) is a binary decision variable for association and ensures that each user can only associate with one base station.

Notice that the above optimisation problem has a binary integer variable, which is constrained to be 0 or 1 (in the objective function and in the constraint functions). Thus it is a typical integer programming problem and characteristically falls into combinatorial optimisation. As a result the problem is NP-complete [122]. Finding the solution of this kind of problem is computationally complex. Even with a very small network setup, finding the optimal solution is time-consuming. For example: one can resort to the *costly* branch-and-bound method, which uses a more suitable upper bound and lower bound to find the exact optimal solution. The branch-and-bound method tacitly enumerates all the feasible solutions, where it partitions all the feasible sets into smaller subsets and then computes certain bounds on the objective function within some of the subsets to eliminate them from other subsets and thus find the optimal solution. For details on the branch-and-bound method, we refer readers to the book by Dimitri Bertsekas [112].

Interestingly the above optimisation problem has only one integer variable. One way of obtaining the solution is to perform constraint relaxation, by which we can let the integer constraint $x_{mi} \in \{0, 1\}$ be $x_{mi} \in [0, 1]$, i.e. the binary constraint is replaced by an interval constraint. Also the Lagrangian relaxation method, which is based on the weak duality theorem, can be applied to solve an integer programming problem [112].

However both relaxation methods can produce a suboptimal solution, i.e. an underestimate of the optimal solution, and it is generally considered as the lower bound of the optimal performance. Note that both constraint relaxation and Lagrangian relaxation can produce the exact same lower bounds if the objective functions and constraint functions in an integer programming optimisation problem are linear [112]. Furthermore, if the optimisation problem has the separability property, which enhances duality, it can be utilised to show that the duality gap tends to diminish. In fact it can be shown that the duality gap diminishes to zero with a high degree of separability. As a result the solution obtained by dual decomposition, which is the core concept of Lagrangian relaxation, turns out to approximately match the optimal primal solution and, more importantly, this method lowers the computational complexity significantly. Therefore we introduce both relaxation methods in the following. The constraint relaxation method provides a centralised solution of the association variable, the bandwidth and the delay. On the other hand, Lagrangian relaxation provides a distributed solution with a lower computational complexity than the centralised solution.

Before applying constraint relaxation, we examine the convexity property of the objective function as given by (4.8a). It should be noted that the objective function in the equivalent GNBS transformation for the proposed cache-association problem is the maximisation of the logarithmic delay function. Thus it preserves concavity as a function of delay. The constraint functions are affine. The rules of convex optimisation therefore apply.

4.4.1 Centralised Solution: Constraint Relaxation Approach

Since the GNBS formulation for a cache-association problem is convex programming, there exists a unique and Pareto-optimal solution. To obtain the centralised solution we reformulate the optimisation problem given by (4.8a)- (4.8f) as follows after performing

constraint relaxation:

$$\underset{\boldsymbol{x},\boldsymbol{w},\boldsymbol{t}}{max} \quad \sum_{m\in\mathcal{M}} \log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) \tag{4.9a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} x_{mi} = 1, m \in \mathcal{M}, \tag{4.9b}$$

$$\sum_{i\in\mathcal{N}_m} t_{mi} \leq t_\tau, m \in \mathcal{M}, \tag{4.9c}$$

$$w_{mi} = \left(A_{mi} \cdot \frac{x_{mi}}{t_{mi}}\right), i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{4.9d}$$

$$\sum_{m\in\mathcal{M}} w_{mi} = W_i^{max}, i \in \mathcal{N}_m, \tag{4.9e}$$

$$x_{mi} \geq 0, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{4.9f}$$

$$x_{mi} \leq 1, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{4.9g}$$

KKT Conditions

The Lagrangian $\mathcal{L}(\cdot)$ is derived in order to find an optimum of the objective function (4.9a), ignoring the constraint (4.9f):

$$\begin{aligned}\mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda},\boldsymbol{\mu}) = &\sum_{m\in\mathcal{M}} \log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) + \sum_{m\in\mathcal{M}} \alpha_m\Big(\sum_{i\in\mathcal{N}_m} x_{mi} - 1\Big) \\ &+ \sum_{m\in\mathcal{M}} \beta_m\Big[\sum_{i\in\mathcal{N}_m} (t_{mi} - t_\tau)\Big] + \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{N}_m} \gamma_{mi}\Big[w_{mi} - \left(A_{mi}\cdot\frac{x_{mi}}{t_{mi}}\right)\Big] + \\ &\sum_{m\in\mathcal{M}}\Big[\sum_{i\in\mathcal{N}_m} \lambda_i(w_{mi} - W_i^{max})\Big] + \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{N}_m} \mu_{mi}(x_{mi} - 1),\end{aligned} \tag{4.10}$$

where $\boldsymbol{\alpha} \succeq 0, \boldsymbol{\beta} \succeq 0, \boldsymbol{\gamma} \succeq 0, \boldsymbol{\lambda} \succeq 0, \boldsymbol{\mu} \succeq 0$, are the Karush-Kuhn-Tucker (KKT) multipliers[2] for the constraints. Using the KKT conditions the complementary slackness are

[2] $\succeq$ denotes element-wise operation.

written as follows:

$$\begin{aligned}\beta_m(t_{mi} - t_\tau) = 0, \forall m \in \mathcal{M}, \forall i \in \mathcal{N}_m,\\ \mu_{mi}(x_{mi} - 1) = 0, \forall m \in \mathcal{M}, \forall i \in \mathcal{N}_m.\end{aligned} \tag{4.11}$$

The first-order necessary and sufficient conditions are:

$$\boldsymbol{\nabla}\mathcal{L}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0. \tag{4.12}$$

Solving for the m-th user:

$$\begin{aligned}&\frac{\partial \mathcal{L}_m}{\partial t_{mi}} = 0\\ \Rightarrow &\frac{-1}{t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}} + \beta_m + \frac{\gamma_{mi} A_{mi} x_{mi}}{t_{mi}^2} = 0.\end{aligned} \tag{4.13}$$

$$\begin{aligned}&\frac{\partial \mathcal{L}_m}{\partial x_{mi}} = 0\\ \Rightarrow &\alpha_m - \frac{\gamma_{mi} A_{mi}}{t_{mi}} + \mu_{mi} = 0\\ &\Rightarrow \alpha_m + \mu_{mi} = \frac{\gamma_{mi} A_{mi}}{t_{mi}}.\end{aligned} \tag{4.14}$$

$$\begin{aligned}&\frac{\partial \mathcal{L}_m}{\partial w_{mi}} = 0\\ &\Rightarrow \gamma_{mi} + \lambda_i = 0\\ &\Rightarrow \gamma_{mi} = -\lambda_i.\end{aligned} \tag{4.15}$$

Substituting (4.14) into (4.13) yields:

$$\Rightarrow \frac{-A_{mi}}{t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}} + \beta_m + (\alpha_m + \mu_{mi})\frac{1}{t_{mi}} = 0. \tag{4.16}$$

Now if the constraint (4.9c) is active, then $\sum_{i\in\mathcal{N}_m}(t_{mi} - t_\tau) = 0$, which makes the optimisation problem infeasible since the objective function (4.9a) approaches negative infinity.

As a result the constraint (4.9c) is inactive, i.e. $\beta_m = 0$. (4.16) becomes:

$$\begin{aligned} &\Rightarrow \frac{-A_{mi}}{t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}} + (\alpha_m + \mu_{mi})\frac{1}{t_{mi}} = 0 \\ &\Rightarrow \frac{A_{mi}}{t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}} = (\alpha_m + \mu_{mi})\frac{1}{t_{mi}} \\ &\Rightarrow t_{mi} = \frac{(\alpha_m + \mu_{mi})\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right)}{A_{mi}} \end{aligned} \tag{4.17}$$

If α_m and μ_{mi} are known, $t^\star_{mi}$ can be computed from the linear equation (4.17).

$$\sum_{i\in\mathcal{N}_m} \frac{w_{mi}t^\star_{mi}}{A_{mi}} = 1 \tag{4.18}$$

To calculate $w^\star_{mi}$, the constraint function (4.9e) and (4.18) can be utilised. Alternatively Algorithm 1 can be used to compute $w^\star_{mi}$. Since we know the lower and upper bounds of the bandwidth of SBS i, which is defined as $[W_i^{min}, W_i^{max}]$, then to find $w^\star_{mi}$ we can apply the bisection method [123] to obtain the following Algorithm 1.

The optimal cache association $x^\star_{mi}$ can be found once the optimal values of $t^\star_{mi}$ and $w^\star_{mi}$ are calculated. In a similar fashion we can calculate $(\boldsymbol{x}^\star, \boldsymbol{w}^\star, \boldsymbol{t}^\star)$ for all users in the HetNet.

Notice that the complexity of such a centralised solution increases as the number of users and SBSs increases. Furthermore, solving for all users' SBS associations in a centralised manner will be computationally inefficient. We therefore resort to Lagrangian relaxation, which enables us to derive a distributed solution, where each user can solve their own optimisation problem *locally* and thereby avoid the burden of a large computational load.

4.4.2 Distributed Solution: Lagrangian Relaxation Approach

In order to solve combinatorial integer programming problems, the Lagrange method was first introduced by Held and Karp in the 1970s to solve the famous *travelling salesman*

Algorithm 1 Bisection algorithm for calculating $w^{\star}_{mi}$.

REQUIRE Lower bound $l_B := W_i^{min}$, upper bound $u_B := W_i^{max}$, t_{mi}, $\epsilon = 0.001$;
ENSURE $w^{\star}_{mi} : f_{mi}(w^{\star}_{mi}) = 0$, $\forall m \in \mathcal{M}$;
Initialise $k = 1$, $W_i^{rem} = W_i^{max}$;
Set $w_{mi}(k) = l_B$, $w_{mi}(k+1) = u_B$;
Find $w_{mi}(k+2) = \frac{w_{mi}(k)+w_{mi}(k+1)}{2}$;
Obtain $f_{mi}(w_{mi}(k))$ and $f_{mi}(w_{mi}(k+1))$ using (4.18);
IF $f_{mi}(w_{mi}(k)) \times f_{mi}(w_{mi}(k+1)) < 0$;
Then
Set $w_{mi}(k+1) = w_{mi}(k+2)$;
ELSE set $w_{mi}(k) = w_{mi}(k+2)$;
END IF
UNTIL $|w_{mi}(k) - w_{mi}(k+1)| < \epsilon$
The m-th user computes $w^{\star}_{mi}$;
Update $W_i^{rem} = W_i^{max} - w^{\star}_{mi}$.
END

problem [124], [125]. Utilising the weak duality, i.e. by dualising the side constraints, the original optimisation problem can be transformed into a Lagrangian problem [126]. This problem introduces a unique property to the optimisation problem called separability, which makes it possible to decompose the original problem into independent subproblems. This feature greatly enhances the computational efficiency. The resultant Lagrangian problem is easy to solve, and solving the problem provides a lower bound on the optimal value of the primal problem. Furthermore, if there exists a strong duality and Slater's condition holds [127], then it can be shown that the duality gap is zero, which means that the primal reaches its minimum as the dual reaches its maximum. Thus the optimal solution of the primal problem can be derived by finding the optimum of its dual problem.

For our case, we are motivated by the separability property of the dual problem, which is achieved by performing Lagrangian relaxation on the optimisation problem (4.8a)-(4.8f). Using the separability property we devise a distributed solution, by which each user in the network can solve their own independent subproblem regarding cache association. The corresponding Lagrangian function for the objective function (4.8a) is:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{m \in \mathcal{M}} \log \left(t_\tau - \sum_{i \in \mathcal{N}_m} t_{mi} \right) + \sum_{m \in \mathcal{M}} \alpha_m \left(\sum_{i \in \mathcal{N}_m} x_{mi} - 1 \right) + \sum_{m \in \mathcal{M}} \left[\sum_{i \in \mathcal{N}_m} \lambda_i (w_{mi} - W_i^{max}) \right]. \tag{4.19}$$

The dual function is:

$$d(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \max_{\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \tag{4.20}$$

After investigating the primal-dual decomposition method from [72], [73], [112] and [122], we can write the dual problem of the primal (4.8a) as:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\lambda}} \quad d(\boldsymbol{\alpha}, \boldsymbol{\lambda}) \tag{4.21a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{N}_m} t_{mi} \leq t_\tau, \tag{4.21b}$$

$$w_{mi} = A_{mi} \cdot \frac{x_{mi}}{t_{mi}}, \tag{4.21c}$$

$$x_{mi} \in [0, 1], \tag{4.21d}$$

$$\boldsymbol{\alpha} \geq 0, \tag{4.21e}$$

$$\boldsymbol{\lambda} \geq 0. \tag{4.21f}$$

The primal problem defined by (4.8a)-(4.8f) has strong duality, since the objective function is a maximisation over a strictly concave function. The constraints are affine and hence log-concave. Therefore a strong duality exists as proven in [128] and [129]. As a result the optimal duality gap is zero. Thus the optimal solution of the primal problem (4.8a)-(4.8f)

can be obtained by solving the dual problem. In fact, when the nodes are large and if the problem is in a separable form, the duality gap diminishes to zero [112]. This phenomenon suits our caching problem, where a large number of users generally will want to associate with SBSs for content-retrieval purposes.

We carefully analyse the Lagrangian function in (4.19) before solving the independent user problem, which can be reorganised as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\lambda}) = \sum_{m\in\mathcal{M}}\left[\log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) + \alpha_m\Big(\sum_{i\in\mathcal{N}_m} x_{mi} - 1\Big)\right] \\ + \sum_{m\in\mathcal{M}}\Big[\sum_{i\in\mathcal{N}_m}\lambda_i(w_{mi} - W_i^{max})\Big] \qquad (4.22)\\ \Rightarrow \mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\lambda}) = \sum_{m\in\mathcal{M}}[F'_m(t_{mi}) + G'_m(x_{mi}) + H'_m(w_{mi})],\end{aligned}$$

where

$$F'_m(t_{mi}) = \log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right), \qquad (4.23)$$

$$G'_m(x_{mi}) = \alpha_m\Big(\sum_{i\in\mathcal{N}_m} x_{mi} - 1\Big), \qquad (4.24)$$

and

$$H'_m(w_{mi}) = \sum_{i\in\mathcal{N}_m}\lambda_i(w_{mi} - W_i^{max}). \qquad (4.25)$$

Now we decompose the dual problem into an independent user problem which can be solved for given $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. The independent user problem is expressed as:

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \mathcal{L}_m(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\lambda}) = &\log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) + \\ &\alpha_m\Big(\sum_{i\in\mathcal{N}_m} x_{mi} - 1\Big) + \sum_{i\in\mathcal{N}_m}\lambda_i(w_{mi} - W_i^{max}) \qquad (4.26)\\ \Rightarrow \min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \mathcal{L}_m(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\lambda}) = &[F'_m(t_{mi}) + G'_m(x_{mi}) + H'_m(w_{mi})].\end{aligned}$$

Each user needs to solve (4.26) to obtain the optimal delay for optimal cache association and optimal bandwidth allocation. Independent user problem can be expressed as follows:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \quad \mathcal{L}_m(\boldsymbol{x},\boldsymbol{w},\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\lambda}) \tag{4.27a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} t_{mi} \leq t_\tau, \tag{4.27b}$$

$$w_{mi} = A_{mi}\cdot\frac{x_{mi}}{t_{mi}}, \tag{4.27c}$$

$$x_{mi} \in [0,1], \tag{4.27d}$$

$$\boldsymbol{\alpha} \geq 0, \tag{4.27e}$$

$$\boldsymbol{\lambda} \geq 0. \tag{4.27f}$$

Here we apply the subgradient method to compute and update the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. In the k-th iteration the dual variables are updated as follows:

$$\alpha_m(k+1) = [\alpha_m(k) + \rho_\alpha(k) g_\alpha(k)]^+, \tag{4.28}$$

where $\alpha_m(k)$ is the k-th iterate, $g_\alpha(k)$ is any subgradient at $\alpha_m(k)$ such that $g_\alpha(k) = \partial_\alpha[\mathcal{L}_m(\boldsymbol{x},\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\lambda})] = (\sum_{i\in\mathcal{N}_m} x_{mi} - 1)$, and $\rho_\alpha(k)$ is an appropriate step size of the k-th iteration. Choosing the step size is crucial for convergence. A detailed discussion about the step size selection can be found in [130]. We follow [111] for choosing the step size, which is defined as $\rho_\alpha(k) = \sigma\frac{(u_{B\alpha}-l_{B\alpha})}{||g_\alpha(k)||^2}$, where $u_{B\alpha}$ and $l_{B\alpha}$ are the upper bound and lower bound respectively, σ is a control parameter. σ is defined as $0 \leq \sigma \leq 2$ and at the beginning $\sigma = 2$. σ is updated to $\sigma/2$ if the upper bound does not decrease after a certain number of iterations. Similarly,

$$\lambda_i(k+1) = [\lambda_i(k) + \rho_\lambda(k) g_\lambda(k)]^+, \tag{4.29}$$

where $\lambda_i(k)$ is the k-th iterate, $g_\lambda(k)$ is any subgradient at $\lambda_i(k)$ such that $g_\lambda(k) = \partial_\lambda[\mathcal{L}_m(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\lambda})] = (w_{mi} - W_i^{max})$, and $\rho_\lambda(k)$ is the step size of the k-th iteration. We again follow [111] for choosing the step size. $\rho_\lambda(k) = \sigma \frac{(u_{B\lambda} - l_{B\lambda})}{||g_\lambda(k)||^2}$, where $u_{B\lambda}$ and $l_{B\lambda}$ are the upper bound and lower bound respectively, σ is a control parameter. The distributed algorithm proposed in this section provides an optimal solution. Though the overall computations are not reduced, the distributed algorithm achieves faster convergence since computations are distributed over the nodes, unlike the centralized algorithm. However the distributed algorithm requires control-message transferring among nodes, which in turn causes large overheads. Therefore the distributed algorithm adopted via Lagrange relaxation may not be suitable for real-time network with a large number of nodes. To avoid this, we propose a heuristic algorithm in the Chapter 5.

4.5 Numerical Results

Numerical results are analysed in this section, obtained by solving the NBG model proposed in Section 4.4. Note that we assume normalised values for parameters for simulation purposes. We will observe how different parameters such as the threshold delay t_τ and the allocated bandwidth per user, affect a user's association decision, i.e. whether it will choose an MBS or an SBS for content retrieval. To validate the proposed NBG model via simulation, a HetNet of two SBSs and a single MBS is considered.

4.5.1 Delay-Function Characteristics

This section analyses the delay function of a user, choosing $C_q = 1$. Instead of being binary, the association variable x_{mi} is allowed to be $x_{mi} \in [0, 1]$. Taking the bandwidth value of each SBS to be 10, the delay function is simulated for a user where the user makes a decision to associate with either of the two SBSs. Fig. 4.4 shows the 3-D plot for a user

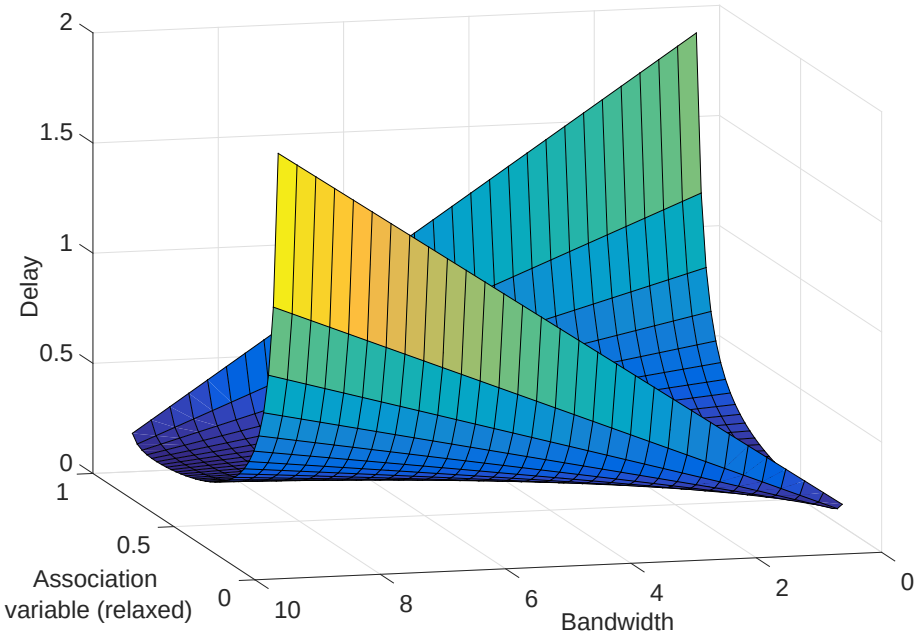


Figure 4.4: Delay function characteristics when bandwidth is 10.

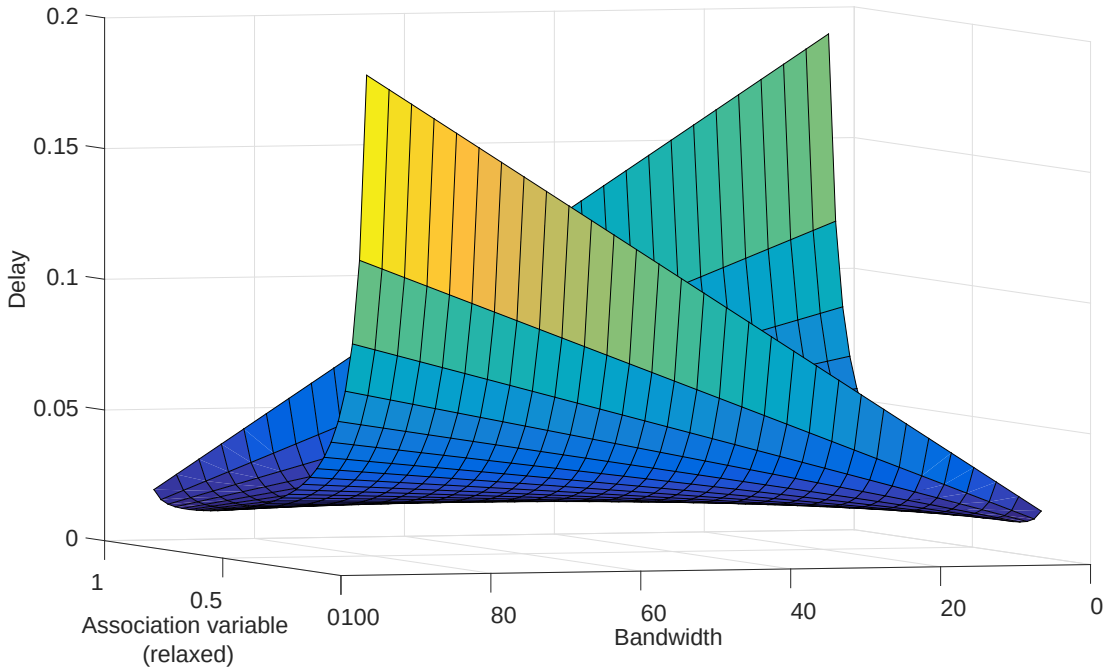


Figure 4.5: Delay function characteristics when bandwidth is 100.

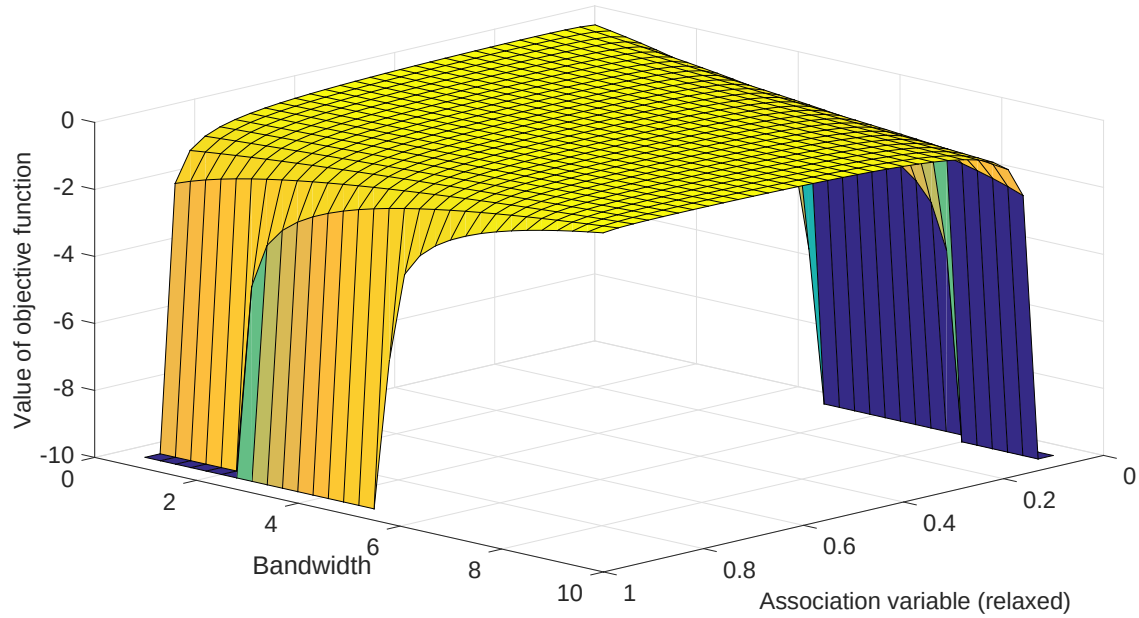


Figure 4.6: Objective function analysis when each SBS bandwidth is 10.

where the delay function is plotted against the association variable and bandwidth. The maximum delay is 2 in this case.The delay function of the user is given by the surface as shown in Fig. 4.4. Fig. 4.5 is obtained for the high-bandwidth case where each SBS bandwidth is chosen to be 100. The maximum delay for a user is 0.02 in this case.

4.5.2 Objective Function Analysis

This section investigates the characteristic objective function proposed in Section 4.4. The objective function for any user m is modified as follows:

$$\log\left(t_\tau - \sum_{i\in\mathcal{N}_m} t_{mi}\right) = \log\left[t_\tau\left(1-\frac{1}{t_\tau}\sum_{i\in\mathcal{N}_m} t_{mi}\right)\right] = \log(t_\tau)+\log\left(1-\frac{1}{t_\tau}\sum_{i\in\mathcal{N}_m} t_{mi}\right) \quad (4.30)$$

The threshold delay t_τ is chosen as 1 and $C_q = 1$. The association variable is allowed to be $x_{mi} \in [0,1]$ and, taking the bandwidth value of each SBS to be 10, the objective function (4.30) is simulated using a 3-D surface plot as shown in Fig. 4.6 for the SBS-association

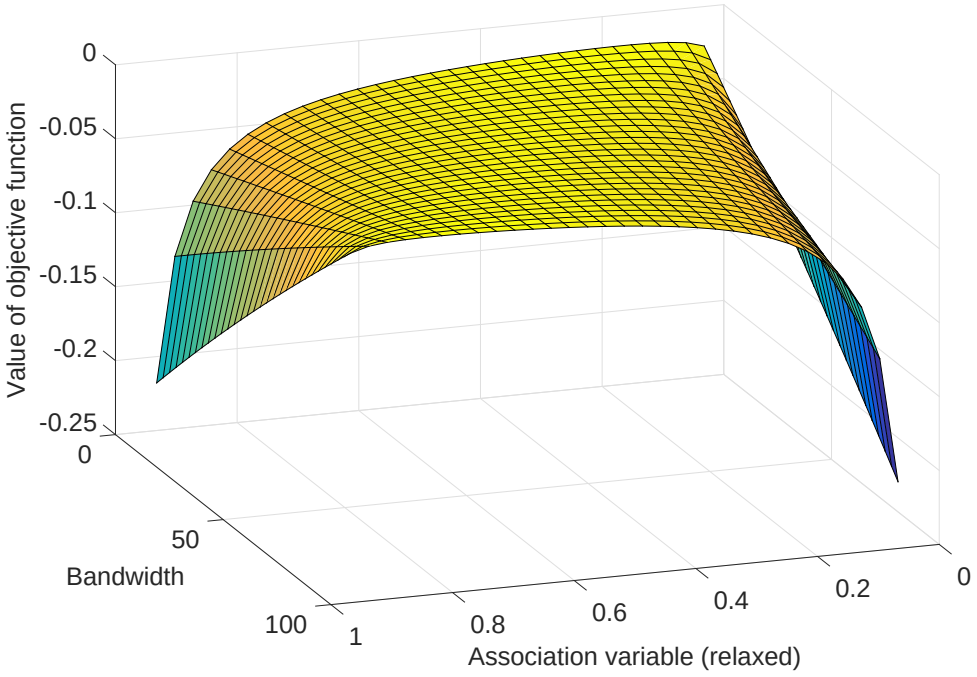


Figure 4.7: Objective function analysis when each SBS bandwidth is 100.

case of user m, where $\mathcal{N}_m = \{1, 2\}$. From Fig. 4.6 it is clear that the surface of the objective function (4.30) is not convex as bandwidth and association variable. However a user can only associate with one SBS, and it has been observed that as long as the conditions, $t_\tau > t_{mi}$ and $t_\tau - t_{mi} > 0$ are satisfied, the objective function is concave with respect to t_{mi}, which is detailed in Fig. 4.8. The highest value of the objective function is -0.2231. Since $\sum_{i\in\mathcal{N}_m} t_{mi}$ is very close to 1 and hence $\log(1 - \sum_{i\in\mathcal{N}_m} t_{mi})$ returns a negative value. Fig. 4.7 is achieved by increasing the bandwidth value of each SBS to 100, while keeping all other parameters unchanged. The shape of the surface of the objective function has changed drastically in this case. Comparing with Fig. 4.6, we can observe that the objective value is increased and the maximum value of the objective function is -0.02 due to the increase in bandwidth.

Fig. 4.8 shows the graphical representation of objective function (4.30) as a function of the user delay. Keeping $t_\tau = 100$ the delay t_{mi} is varied and the value of the objective

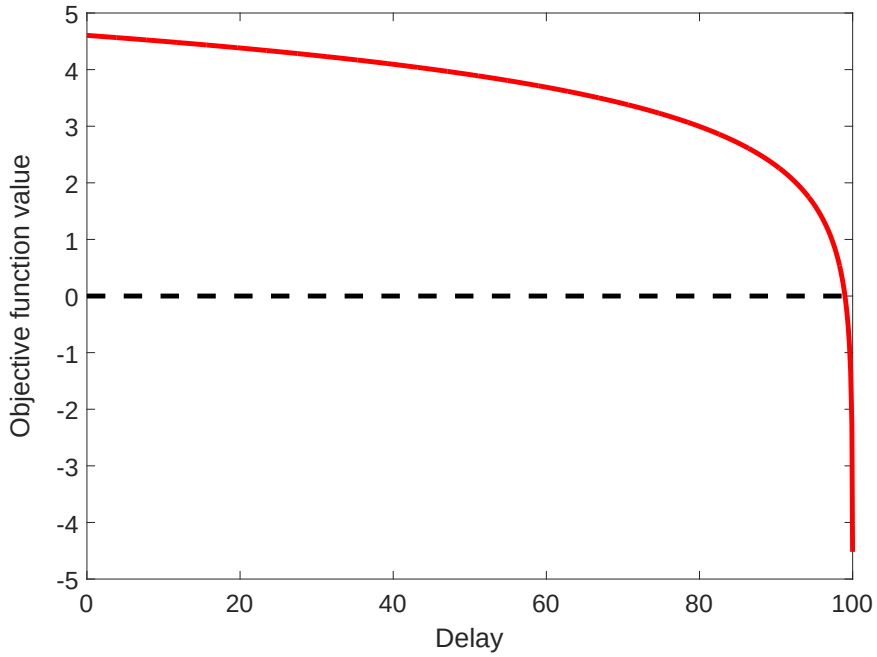


Figure 4.8: Concavity of the objective function as a function of delay.

function is plotted in the y-axis. It shows that the objective function follows concavity. Notice that when the delay $t_{mi} = 99$ the objective function value becomes 0 and becomes negative for any $t_{mi} > 99$. In an NBG, each user tries to maximise its objective function via cooperation. In this case, the interpretation of any negative objective value is that a user's threshold delay requirement is barely satisfied.

4.5.3 Equilibrium Performance Analysis

This section provides a quantitative analysis of the content download delay performance of the proposed cache-association scheme. Exhaustive search is a well-known method to compute the solution of an integer programming problem [131]. For a large system exhaustive search requires an increasing computation time. However it has been used in several literatures to compute the optimal solution of an optimisation problem [73], [132], [133], [134], [135], [136] due to its implementation simplicity and often compared with

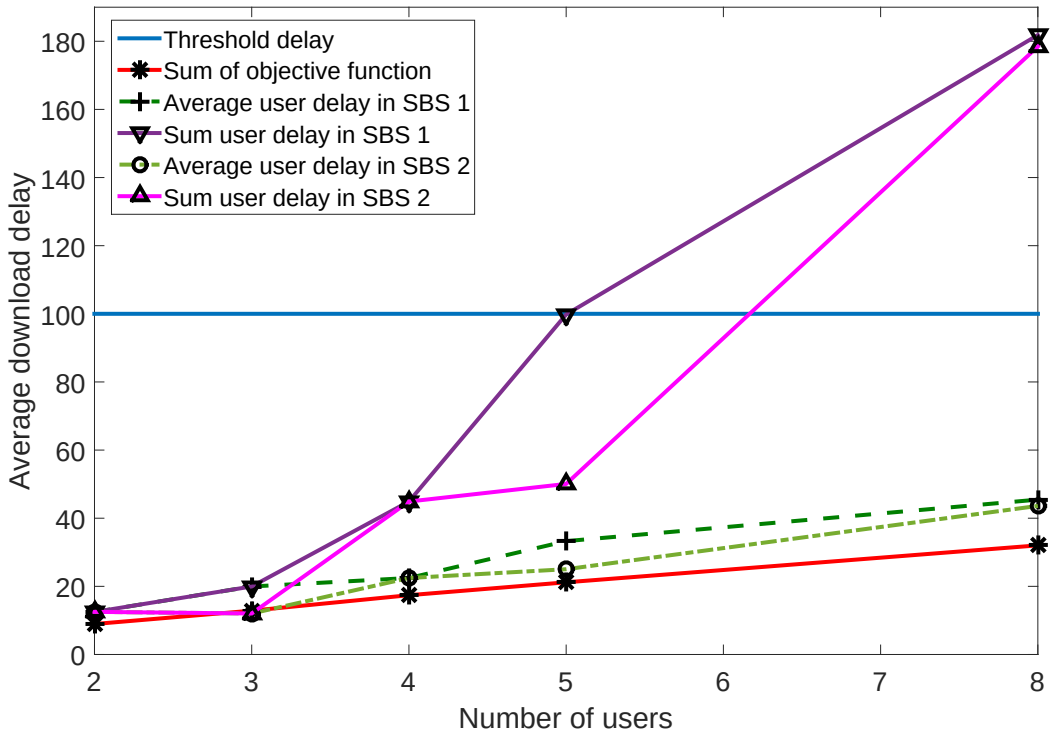


Figure 4.9: Quantitative analysis of proposed cache-association scheme.

the performance of various suboptimal algorithms. The major drawback is the increase in the computational overhead for a large number of nodes. However, there are various solvers such as genetic algorithm (GA) solver in MATLAB or a commercial solver such as GAMS can be used to analyse the performance of the proposed algorithm for a bigger subsystem. To overcome the computational inefficiency of the proposed algorithm, we propose a heuristic solution in Chapter 5.

Here a MATLAB simulation is performed using exhaustive search method by varying the number of users in the HetNet, while the threshold delay t_τ and the SBS-bandwidths are kept fixed. For different user settings, we numerically compute the optimal association and bandwidth allocation. The numerical values of the parameters are: $W_1^{max} = 10$, $W_2^{max} = 10$, $\Upsilon_1 = \Upsilon_2 = 10$, $C_q = 100$, $t_\tau = 100$, and $M = 14$. We will compare the

individual user delay and average user delay with the delay threshold. An individual user delay is found by observing the equilibrium point for each user setting. Thus for each user setting we can also calculate an average user delay using the data from individual user delays. Finally, we will validate that in every user setting the threshold delay is satisfied for both the individual and average user-delay cases. In addition, we will also calculate the sum delay for each user settings and observe when it exceeds the threshold delay.

We compute the average delay per user and the value of the objective function for each SBS association and plot them against different user-settings as shown in Fig. 4.9. The MATLAB simulation provides numerical values of the user association and bandwidth allocation for each network configuration. Using these data each user's delay is calculated. Finally the value of the objective function for each user is calculated for different network settings and plotted as the red line in Fig. 4.9. As expected the sum objective function value increases when the number of users increases in the HetNet.

We have also calculated and plotted the average user delay at each SBS to compare the SBS-specific delay performance. When an odd number of users are present in the HetNet, the average download delay for users in SBS 1 increases, as seen for the 3-user, 5-user cases in Fig. 4.9 given by the dark-green dotted line. The NBG algorithm always assigns more users to SBS 1 in each odd combination of user settings since both the association constraint and the bandwidth-allocation constraint are satisfied. For the 3-user case the the assigned bandwidths for 2 users for SBS 1 association are 4.98 and 5.01 respectively and the average delay is 20.2651. For the combination of even number of users in the HetNet, both SBSs provide identical performance in terms of association and bandwidth allocation. It is observed that the average download delay performance of both SBSs, plotted by the two green lines in Fig. 4.9, satisfy the threshold delay requirement, which is plotted by a blue line. The sum-delay in both SBSs exceeds the threshold delay as plotted by the violet and pink lines respectively for SBS 1 and SBS 2 in Fig. 4.9. However

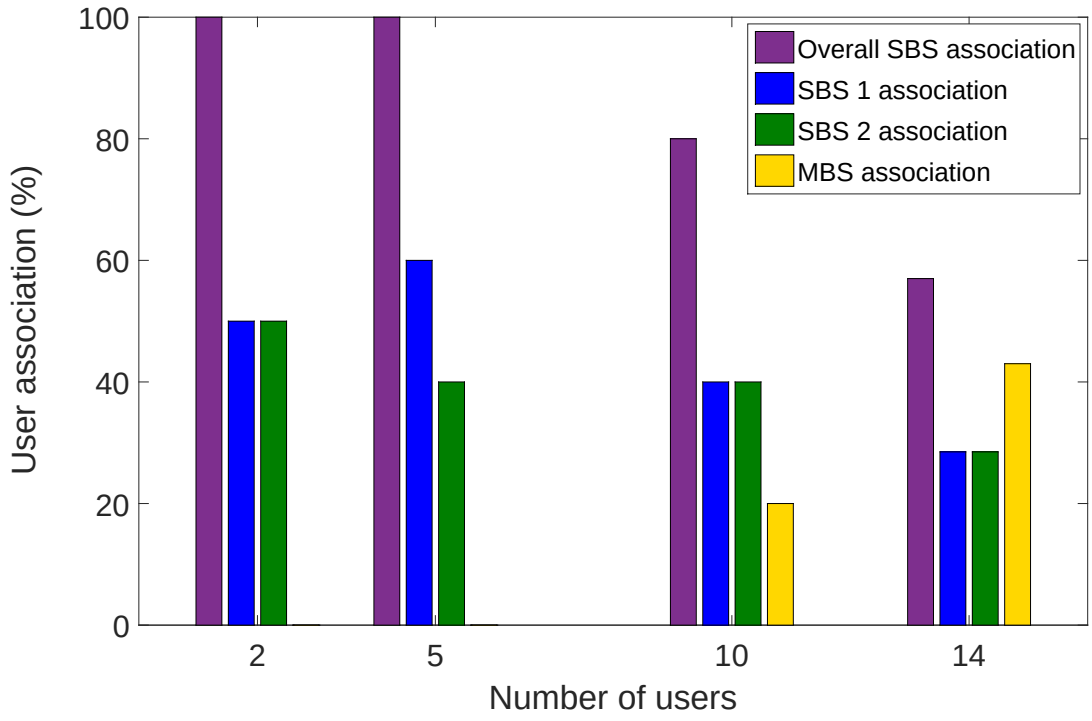


Figure 4.10: Effect of flexible threshold delay on the SBS association.

this is completely fine since the individual user delay does not exceed the threshold delay requirement. Note that after accommodating 8 users in two SBSs, the NBG-based cache-association scheme does not allow any more users to associate with either of the SBSs since the remaining bandwidth is not enough to fulfil the threshold-delay requirement.

4.5.4 Impact of Threshold Delay in SBS Association

Here we investigate the impact of the threshold delay for both user association and bandwidth allocation. The goal is to investigate the network dynamics, i) when the delay requirement is flexible and ii) when the delay requirement is stringent. It is assumed that both SBSs have the same bandwidth and they have identical SNR performance. The values of the parameters used in the simulation are: $W_1^{max} = 10$, $W_2^{max} = 10$, $\Upsilon_1 = \Upsilon_2 = 10$, $C_q = 100$, $t_\tau = 100$, and $M = 14$.

Flexible Delay

We vary the number of users and note the allocated bandwidth and delay for each user. For different user settings we calculate the % user association which is plotted against the number of users as shown in Fig. 4.10. % user association captures the efficiency of the proposed NBG-based cache-association scheme and provides a trade-off between bandwidth allocation and the delay parameter. Most importantly the threshold delay allows an MNO to have optimum association at each SBS. % user association is defined as follows which actually captures the efficiency of the caching system:

$$\text{User association in\%} = \frac{\text{Associated users to a particular SBS or an MBS}}{\text{Total users participating in the bargaining}} \times 100. \quad (4.31)$$

It is observed that, when an even number of users are present in the network, they are distributed to each SBS equally. However when five users are present in the HetNet, three of them choose SBS 1 and the remaining two users choose SBS 2. This is due to the fact that all the constraints for user 3 are satisfied for associating with SBS 1. By keeping the minimum bandwidth allocation per user to 2, it is observed that the SBSs can accommodate up to 8 users for the given bandwidth and threshold-delay settings. For the simulation settings of 10 users, 2 users associate with the MBS, and with 14 users in the HetNet, 6 users associate with the MBS. For user settings of 2 and 5, 100% of users will associate with SBSs. For the user settings of 10 and 14, the SBS association is 80% and 57% respectively. Thus the MBS will have to serve 20% and 43% of the content download requests respectively in those cases. This means that those users who are associating with an MBS have failed to successfully negotiate their cache-association SBS. It is analogous to the case when an SBS does not have enough bandwidth to serve all incoming content-download requests. Thus it gives an idea of how many users can be served by an SBS, which depends on the threshold delay requirement and the available bandwidth of each

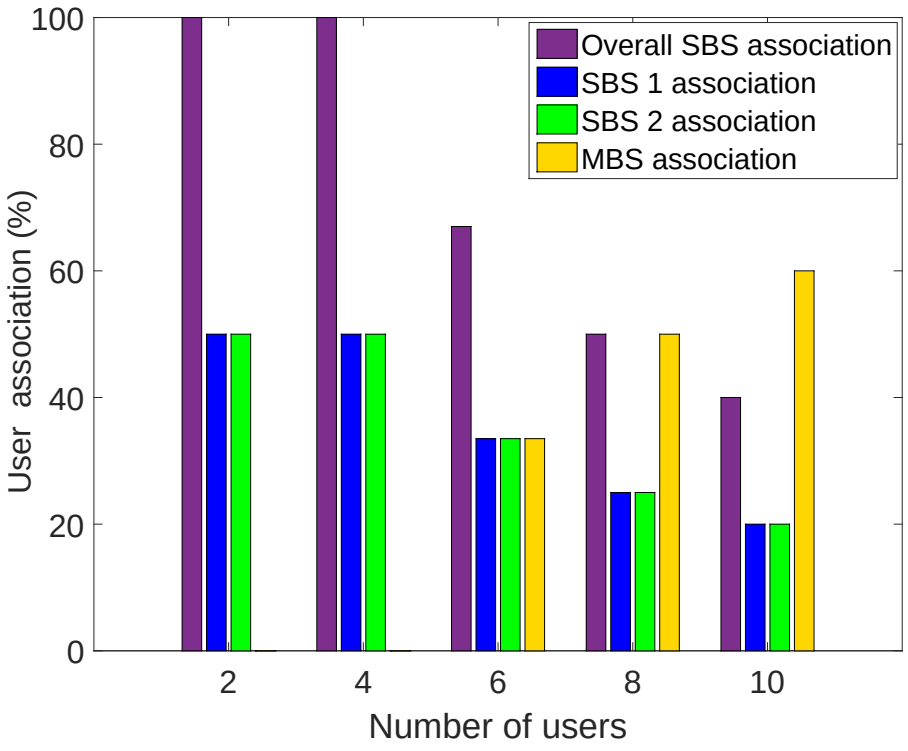


Figure 4.11: Effect of stringent threshold delay on the SBS association.

SBS.

Stringent Delay

The threshold delay requirement is reduced to 75% compared to the case analysed in Section 4.5.4. The value of t_τ is changed to 25 while all other parametric values remain unchanged as given in Section 4.5.4. A reduction in the threshold delay means that each SBS will have to allocate more bandwidth to users to satisfy the delay constraint, and as a consequence less number of users can be served. The network selection dynamics is observed for 5 different user settings as shown in Fig. 4.11. It is found that a maximum of two users can associate with each SBS, while the remaining users join the MBS to download their desired contents. Thus in the 2-user case and in the 4-user case, the

MBS does not have to share the content-caching burden and all users select one of the SBSs. However, in the 6-user, 8-user and 10-user cases, the SBSs' content-caching load remains unchanged, since each SBS has already allocated almost all of their bandwidth to the users. As a result there is not enough bandwidth remaining to allocate to any new arrivals satisfying the threshold delay requirement.

Remark 1. *An important observation is that an MNO can utilise the threshold delay t_τ as a QoS parameter to increase or decrease the MBS's content-caching load. A flexible delay requirement will certainly increase the SBS association for a given system bandwidth. Thus a flexible threshold delay is suitable for those wireless services which can withstand delay variations. On the other hand a stringent threshold delay can be implemented with the dedicated SBSs, which will provide delay-constrained contents only. Most importantly the threshold delay-based content-caching system can be utilised to control the MBS' content-caching burden.*

4.5.5 Threshold Delay Impact on Average Delay Performance

Fig. 4.12 shows the average content download delay performance of users subjected to different threshold delay requirements. In the studies conducted, t_τ is varied and we observe the change in the average delay performance of users while the SBS bandwidth remains unchanged. As shown in Fig. 4.12 for $t_\tau = 100$, the two SBS setups can accommodate up to 8 users and the average delay per user increases, which is given by the red line. When 8 users associate with the two SBSs the average content download delay per user is 45. When more than 10 users are present in the HetNet, the NBG-based cache-association algorithm does not accept any more users, since the bandwidth allocation of a user goes below 1. Also the individual delay becomes greater than 100, which violates the threshold delay constraint. When $t_\tau = 25$, more bandwidth per user needs to be allocated to satisfy

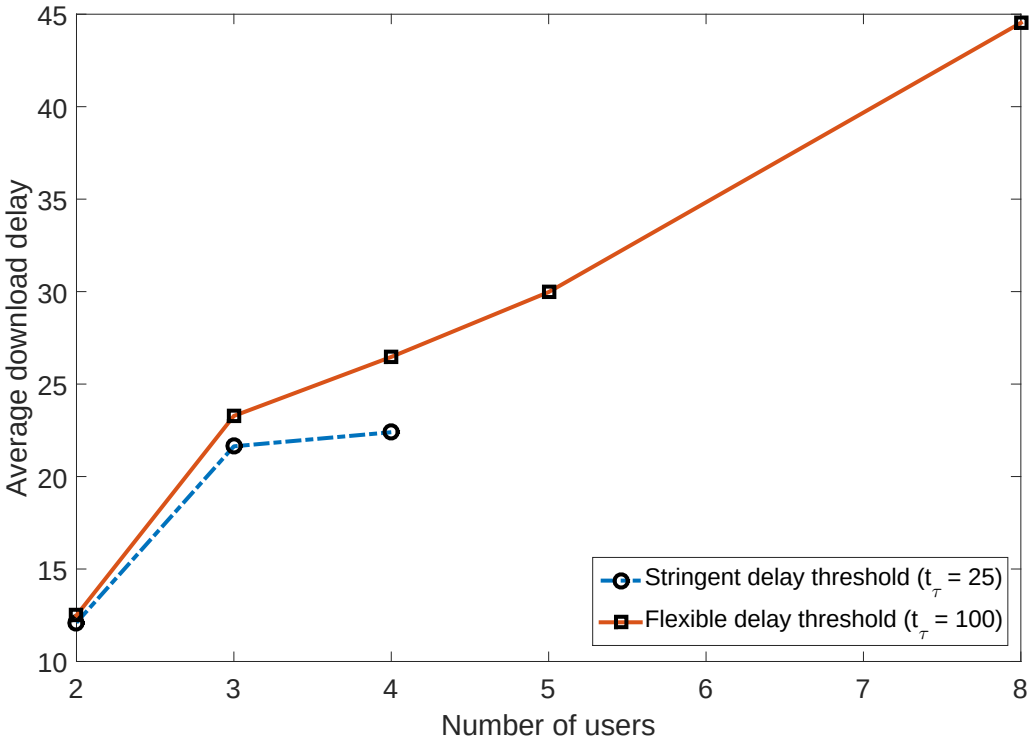


Figure 4.12: Effect of different threshold delays on average delay performance.

the delay requirement. Thus only two users per SBS are accepted by the NBG-based cache-association algorithm, which is given by the blue line.

4.6 Summary

In this chapter, a novel NBG-based content-caching system is proposed for a HetNet with a view to maximising the SBS association. The average content-retrieval delay from an MBS is considered as the threshold delay, which an SBS must satisfy to accommodate users while delivering content. A unique yet simple utility function is devised and then formulated into a cooperative bargaining game. A centralised solution and a distributed solution are proposed for SBS-based content association-retrieval systems using constraint relaxation and Lagrange relaxation respectively. Table 4.2 gives a comparison between

Table 4.2: A comparison between constraint relaxation and Lagrange relaxation.

No.	Properties	Constraint relaxation	Lagrange relaxation
1	Time complexity	Depends on solution methodology. For example, with branch and bound method time complexity is $\mathcal{O}(N)$; with exhaustive search time complexity is exponential.	Solvable in polynomial time.
2	Algorithm type	Centralised.	Distributed.
3	Parallel computation	No.	Yes.

the two methods. The inclusion of the threshold delay into the utility function design allows us to replicate a realistic content-caching scenario, and it has been shown that the users will associate with an SBS when the association requirement is satisfied, given that the content is available in the small-cell tier. A trade-off between threshold delay and a user's allocated bandwidth is also established, by utilising which an SBS can fine-tune its content-caching delay as well as have optimum user association.

Chapter 5

Heuristic Approach for Cache Association and Bandwidth Allocation

5.1 Chapter Introduction

This chapter presents a heuristic approach to solve the cache-association and bandwidth-allocation problem in a 5G HetNet. The distributed algorithm proposed in Chapter 4 is based on the subgradient method, which comes with high computational overheads. The joint optimisation problem proposed in Chapter 4 is first decoupled into a user-association problem and a bandwidth-allocation problem where users are assigned to an SBS first and then the bandwidth allocation for each user is performed for the corresponding SBS. One of the benefits of the heuristic algorithm is that it does not require global knowledge regarding the network topology and provides an approximate solution which is very close to the original solution [137], [138]. In this chapter a heuristic algorithm is developed for cache association and bandwidth allocation for users, emphasising distributed and

decentralised implementation.

5.2 Network Model

Assume a two-tier HetNet-based content-caching system, where the MNO deploys one MBS and a large number of SBSs. The MBS and SBSs adopt LTE in their physical layer to provide services to users. The MBS and SBSs are connected with the core network with a high capacity, and a limited-capacity wired backhaul respectively. The set of all SBSs is denoted by $\mathcal{I} = \{1, \cdots, I\}$, where $i \in \mathcal{I}$ denotes the i-th SBS. We denote the set of users as $\mathcal{M} = \{1, \cdots, M\}$, where $m \in \mathcal{M}$ denotes the m-th user. Assume that there exists a global content library defined by the set $\mathcal{Q} = \{1, \cdots, Q\}$. The MNO deploys SBSs very close to each other so that they can serve as many content requests as possible. It is assumed that the MBS and SBSs are deployed in an underlay spectrum-access manner and each SBS has the same SNR, which is represented by Υ. A delay-based utility function is considered and is given by the following function:

$$t_{mi} = \frac{x_{mi} C_q}{w_{mi} \log(1 + \Upsilon)}, \tag{5.1}$$

where x_{mi} is the binary association variable, C_q is the size of the q-th content and w_{mi} is the allocated bandwidth for the m-th user for associating with the i-th SBS. For each user the minimum delay requirement is t_τ.

5.3 The Heuristic Fair Caching Algorithm

This section details the heuristic algorithm to solve the cache-association problem presented in Chapter 4. The proposed heuristic algorithm is called the *heuristic fair caching* (HFC) algorithm. It aims to emulate the fair resource allocation characteristics of an NBG-based caching system. The HFC algorithm differs from the NBG-based joint cache-

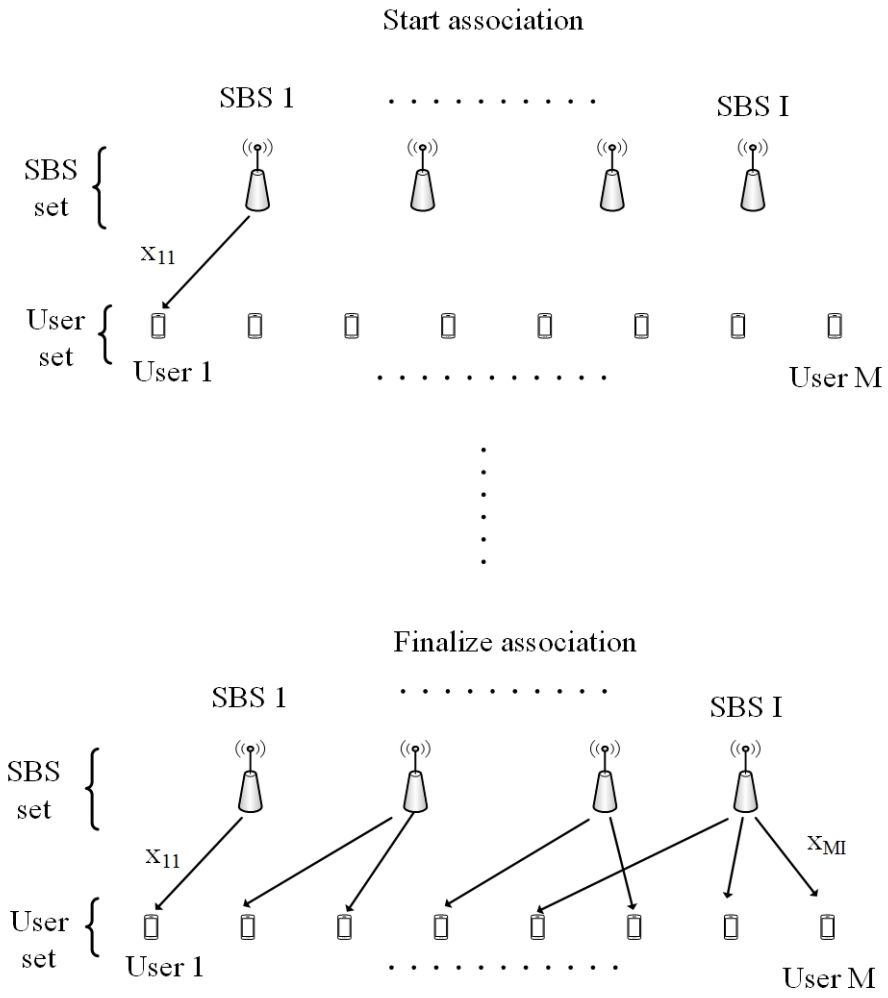


Figure 5.1: Illustration of user association in an SBS-based caching system using the heuristic approach.

association and bandwidth allocation algorithm as presented in Chapter 4, in that the HFC algorithm will determine the cache association first. It is assumed that the minimum required bandwidths for all users are the same and they are assigned to SBSs sequentially. In the second phase, the HFC algorithm performs bandwidth allocation to improve the delay requirement further. Thus it requires two stages to execute its operation. In the following both stages are clearly described.

5.3.1 Association Stage

Fig. 5.1 details the cache-association stage, where each user is assigned to an SBS sequentially. The HFC algorithm first picks any user m and assigns it to SBS i assuming that all users require the same bandwidth. This bandwidth is pre-calculated so that a minimum threshold delay condition is always satisfied. After that another user is picked for SBS i as long as the bandwidth criterion is satisfied. Fig. 5.1 shows the final condition of the association stage, where SBSs have a different numbers of users. This is due to the fact that a user must remain under the coverage of an SBS for association. In addition each user's delay requirement must be satisfied. For example: the I-th SBS has three users to serve and SBS 1 has only one user to serve. This orientation reflects a realistic scenario where a user may not exist under the coverage of multiple SBSs or the case where SBSs have a limited bandwidth. Once the association stage is completed, the HFC algorithm starts the bandwidth-allocation stage.

5.3.2 Bandwidth-Allocation Stage

Fig. 5.2 illustrates the bandwidth-allocation procedure. Since the association is fixed, the bandwidth allocation of a particular SBS to its users can be finalised. Each SBS may need to serve more than one user. In that case the bandwidth will be proportionally divided among users. Fig. 5.2 shows such a scenario, where the I-th SBS needs to serve at least 3 users. On the other hand user 1 is associated with SBS 1 and does not need to share the SBS bandwidth with any other users.

The HFC algorithm follows a *waterfilling-like* bandwidth-allocation pattern for the users associated with a particular SBS, except that each user receives some non-zero bandwidth allocation. In this case the assumption of equal bandwidth allocation is replaced by a constraint in the modified NBG formulation, where each user requires a certain

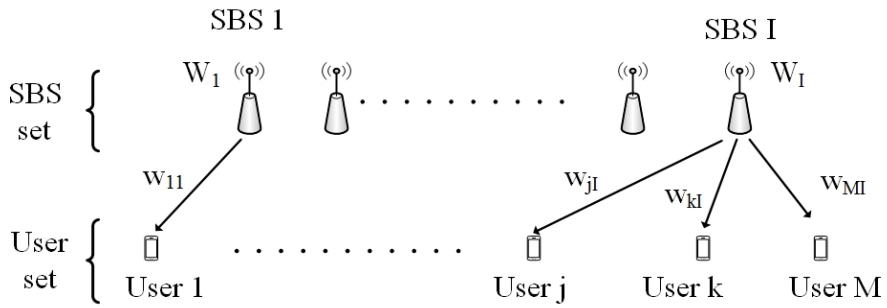


Figure 5.2: Illustration of bandwidth allocation in an SBS-based caching system using the heuristic approach.

amount of bandwidth depending upon its content size and threshold delay requirement. Thus the HFC algorithm finalises the bandwidth allocation for all users who associate with an SBS for content-retrieval purposes. Since the algorithm runs in each SBS, only local knowledge of the network resources is required.

Now recall the NBG formulation in Section 4.4 of Chapter 4. We can modify the NBG by fixing the SBS association for each user, and the optimisation problem takes the following form:

$$\max_{\boldsymbol{W}} \quad \sum_{m\in\mathcal{M}} \left[\log t_\tau - \log(1-\frac{a_m}{w_m})\right], \tag{5.2a}$$

$$\text{s.t.} \quad \sum_{m\in\mathcal{M}} w_m = W^{max}, \forall m \in \mathcal{M}, \tag{5.2b}$$

$$w_m > a_m, \forall m \in \mathcal{M}, \tag{5.2c}$$

where $a_m = \frac{C_q}{t_\tau \log(1+\Upsilon)}$ is an arbitrary constant for the m-th user. The objective function (5.2a) is a concave function. The constraint functions (5.2b) and (5.2c) are affine. Therefore the optimisation problem given by (5.2a)-(5.2c) is a convex optimisation. In addition, the optimisation problem satisfies all the conditions of an NBG. Indeed the optimisation problem is a simplified version of the NBG-based optimisation proposed in Chapter 4,

with reduced computational complexity since the user association is fixed. As a result a unique optimal solution can be found for the problem (5.2a)-(5.2c). Furthermore the logarithmic utility function guarantees fairness in bandwidth allocation [86], [88]. In the following the KKT optimality conditions are derived.

KKT Conditions

The Lagrangian $\mathcal{L}(\cdot)$ is derived as follows to find an optimum of the objective function (5.2a) :

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{m \in \mathcal{M}} \left[\log t_\tau - \log \left(1 - \frac{a_m}{w_m} \right) \right] + \lambda \left(\sum_{m \in \mathcal{M}} w_m - W^{max} \right) - \sum_{m \in \mathcal{M}} \mu_m (a_m - w_m), \tag{5.3}$$

where $\boldsymbol{\lambda} \succeq 0, \boldsymbol{\mu} \succeq 0$ are the nonnegative KKT multipliers for the constraints. Using the KKT conditions the complementary slackness is written as follows:

$$\mu_m (a_m - w_m) = 0, \forall m \in \mathcal{M}. \tag{5.4}$$

The first-order necessary and sufficient conditions are:

$$\boldsymbol{\nabla} \mathcal{L}(\boldsymbol{W}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0. \tag{5.5}$$

Solving for the m-th user:

$$\begin{aligned} \frac{\partial \mathcal{L}_m}{\partial w_m} &= 0 \\ \Rightarrow \frac{1}{w_m - a_m} - \frac{1}{w_m} - \mu_m &= -\lambda \end{aligned} \tag{5.6}$$

Now if the constraint (5.2c) is active then $w_m = 0$ which makes the bandwidth allocation problem infeasible since the objective function approaches negative infinity. Therefore w_m should be positive, and by analysing the objective function (5.2a) we can say that

$w_m > a_m$. As a result the constraint (5.2c) is inactive and $\mu_m = 0$, and (5.6) becomes:

$$\Rightarrow \frac{1}{w_m - a_m} - \frac{1}{w_m} = -\lambda \tag{5.7}$$

Therefore the following equality condition is true:

$$\frac{w_1(w_1 - a_1)}{a_1} = \cdots = \frac{w_j(w_j - a_j)}{a_j} = \frac{w_k(w_k - a_k)}{a_k} = \cdots = \frac{w_M(w_M - a_M)}{a_M} = -\lambda, \tag{5.8}$$

where the users $j, k \in \mathcal{M}$. An interesting observation is that increasing the bandwidth for any user will result in an increase in the bandwidth allocation of other users associating with the same SBS. For example: if w_1 is increased for user 1 to improve its delay, then according to (5.8) for every user j, their individual bandwidth allocation must be increased to satisfy (5.8). Thus (5.8) exhibits a waterfilling-like characteristic. Using equation (5.8), we can analyse a two-user bandwidth-allocation scenario which explains the waterfilling-like bandwidth-allocation phenomenon more clearly. For the users j and k we can write the following equality:

$$\frac{w_j(w_j - a_j)}{a_j} = \frac{w_k(w_k - a_k)}{a_k} \Rightarrow \frac{(w_j)^2}{a_j} - w_j = \frac{(w_k)^2}{a_k} - w_k. \tag{5.9}$$

By carefully analysing (5.9) it can be concluded that each user's bandwidth function is parabolic in nature. Thus it is possible to sketch the bandwidth function for each user as shown in Fig. 5.3. The bandwidth of each user is plotted on the x-axis and the function of bandwidth given by (5.9) for users j and k is plotted on the y-axis. The black and red curves represent the functions of the bandwidths of user j and user k respectively. Note that the amount of bandwidth that can be allocated to each user depends on the numerical value of a_j for the j-th user, since the allocated bandwidth has to be bigger than a_j, i.e. $w_j > a_j$ has to be satisfied. Thus not only does user k's bandwidth allocation depend on the numerical value of a_k, where $w_k > a_k$, but also the condition $W^{max} - w_j > a_k$ needs to be satisfied such that the feasible bound becomes $a_k < w_k \leq W^{max} - w_j$. In other

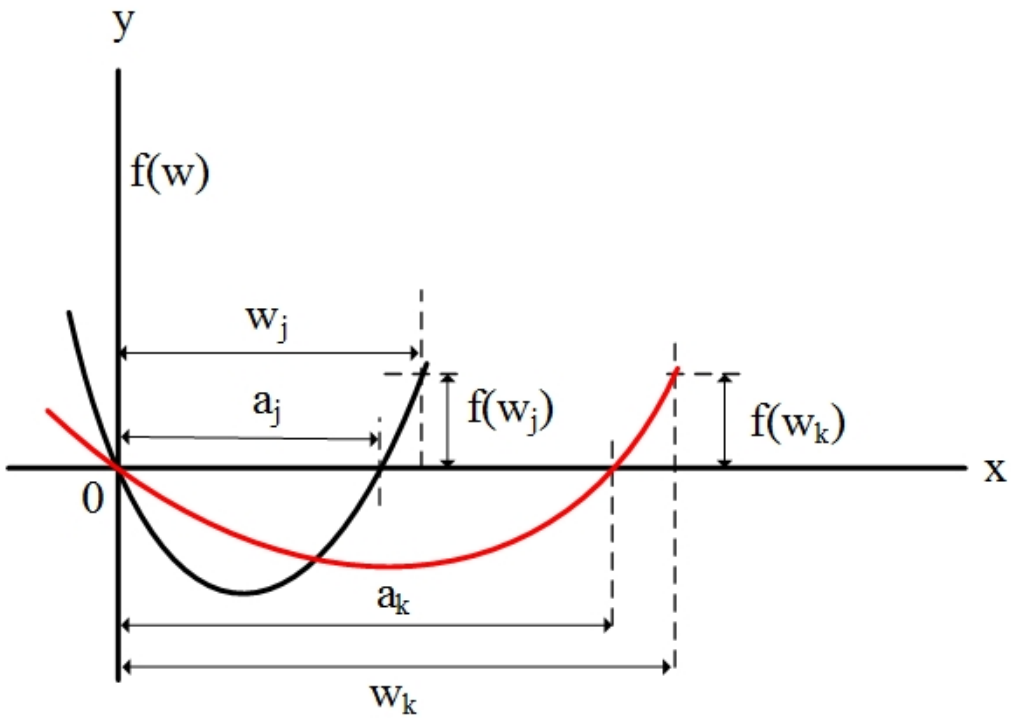


Figure 5.3: Analysis of the bandwidth-allocation characteristics for a two-user association case using the HFC algorithm.

words, the remaining bandwidth of a particular SBS must be bigger than the required bandwidth of user k, after it allocates bandwidth to user j.

Notice that if a bargaining is successful among multiple users for sharing SBS resources such as its bandwidth, then the bandwidth requirement for all users will always be satisfied. Otherwise a user will seek to associate with another SBS. In any case if a negotiation is not successful with an available SBS, i.e. the allocated bandwidth is not enough to satisfy the threshold delay requirement of a user, then the user will associate with an MBS. However these associations will be sorted out in the user-association stage. In summary, how many users can associate with a particular SBS depends on its maximum bandwidth. Utilising this concept, we can extend bandwidth allocation to where more than two users want to associate with an SBS.

Using (5.8) the bandwidth-allocation procedure can now be analysed for the multi-user case. From (5.8) the following expression can be obtained:

$$\sum_{m\in\mathcal{M}} \frac{(w_m)^2}{a_m} - \sum_{m\in\mathcal{M}} w_m = \sum_{m\in\mathcal{M}} f_m(w_m) - W^{max}, \tag{5.10}$$

where $f_m(w_m) = \frac{(w_m)^2}{a_m}$. Thus the bandwidth allocation to any user m depends on the value a_m. Since $w_m > a_m$, for a given W^{max}, the number of users that can be admitted to an SBS depends on a_m, $\forall m \in \mathcal{M}$. If the numerical value of each a_m, $\forall m \in \mathcal{M}$, is small then the SBS can serve more users than when each a_m is large. Suppose that an SBS allocates all of its bandwidth to users, then (5.10) becomes:

$$\sum_{m\in\mathcal{M}} f_m(w_m) = W^{max}. \tag{5.11}$$

Thus (5.11) explicitly restricts the number of users who can associate with a particular SBS, depending upon how much bandwidth an SBS has, i.e. on W^{max}. The final step is to prove the convergence of (5.11).

Convergence

In order to prove the convergence of the bandwidth-allocation stage of the HFC algorithm, the first step is to compute the gradient of the bandwidth allocation function. We can graphically represent (5.11) as shown in Fig. 5.4. The left-hand side of (5.11) is plotted by the green line. Similarly the red line plots the bandwidth allocation function of any user m. The x-axis represents the bandwidth and in the y-axis is the function of the bandwidth. The goal is to calculate the gradient of the line segment AB, where the line AB passes through the points $A(x_n, y_n)$ and $B(x_{n+1}, y_{n+1})$ and is a tangent line to the function $f(x_n)$ at the point $A(x_n, y_n)$. Thus the gradient is:

$$f'(x_n) = \frac{y_{n+1} - y_n}{x_{n+1} - x_n}, n = 0, 1, 2, \cdots \tag{5.12}$$

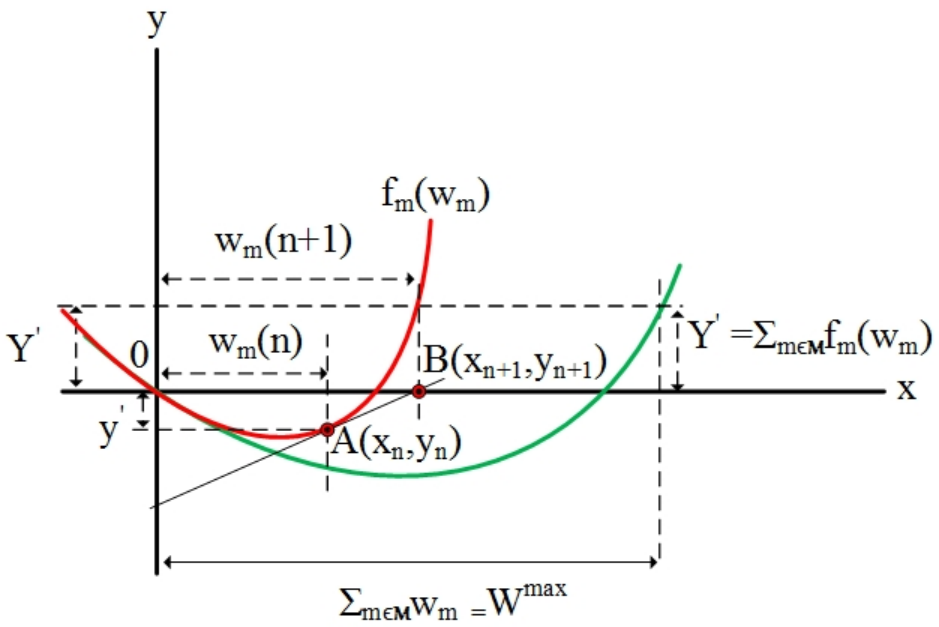


Figure 5.4: Graphical illustration of the convergence properties of the bandwidth allocation function.

By substituting the values of x_n,x_{n+1}, y_n and y_{n+1} (follow Fig. 5.4) we get:

$$f'_m(w_m(n)) = \frac{0 - y'}{w_m(n+1) - w_m(n)}, n = 0, 1, 2, \cdots \tag{5.13}$$

Now $y' = y_n = f(x_n) = f_m(w_m(n))$. Equation (5.13) becomes:

$$f'_m(w_m(n)) = \frac{f_m(w_m(n))}{w_m(n) - w_m(n+1)}, n = 0, 1, 2, \cdots \tag{5.14}$$

From (5.14) we can write that

$$w_m(n+1) = w_m(n) - \frac{f_m(w_m(n))}{f'_m(w_m(n))}, n = 0, 1, 2, \cdots \tag{5.15}$$

We can see that (5.15) resembles the well-known root-finding method called Newton's method [139], [140]. Newton's algorithm, or Newton's method, is widely used because of its fast-convergence property. In fact the method is quadratically convergent [139] and in some special cases linearly convergent [140]. As a result the number of decimal places of

Algorithm 2 Iterative bandwidth allocation using Newton's method .

1: **REQUIRE** $w_m(0) = w_0$, $\forall m \in \mathcal{M}$ and $\epsilon = 0.001$;

2: **ENSURE** $f_m(w_m) = 0$, $\forall m \in \mathcal{M}$;

3: Initialise $n = 1$;

4: **COMPUTE** $f_m(w_m(0))$ and $f'_m(w_m(0))$, $\forall m \in \mathcal{M}$;

5: **REPEAT**

6: Set $w_m(n) = w_0$;

7: Set $w_m(n+1) = w_m(n) - \frac{f_m(w_m(n))}{f'_m(w_m(n))}$;

8: **UNTIL** $|w_m(n+1) - w_m(n)| < \epsilon$;

9: Update n = n+1.

10: **END IF**

Algorithm 3 HFC algorithm.

1: Set $t_{mi} = 0$ and $\mathcal{M} = \phi$, $\forall m \in \mathcal{M}$;

2: **FOR** $m = 1$ to M {

i Find $i \in \mathcal{N}_m$ satisfying $w_{mi} > 0$ and $t_{mi} < t_\tau$;

ii Set $x_{mi} = 1$;

iii Update $t_{mi} = t_m$ and the user set $\mathcal{M} : \mathcal{M} = \mathcal{M} - \{m\}$. }

3: **WHILE**{

i Get $w_m^\star$ using Algorithm 2;

ii Check feasibility using equation (5.10) and (5.8);

iii Verify $t_m < t_\tau$ for all m;

iv Update t_m with $w_m^\star$ for all users and put $\mathcal{M} = \mathcal{M} - \{m\}$.}

4: **END**

accuracy approximately doubles at each iteration. The only drawback of this algorithm is that it may converge to a different root from the expected root or even diverge, if the initial value is not close enough to the root. The computational complexity of Newton's method is $\mathcal{O}(N \log F(N))$, where $F(N)$ is the cost of computing $\frac{f_m(w_m(n))}{f'_m(w_m(n))}$ with N digit accuracy. We develop an iterative algorithm based on Newton's method to calculate the bandwidth of the users associating with SBSs in a distributed manner, given by Algorithm 2.

Thus the proposed HFC algorithm executes after finishing two stages. Algorithm 3 summarises the two stages from initiation to execution.

Remark 2 (Residual utility)**.** *According to the objective function (5.2a), a user may have a residual utility value equivalent to its threshold delay without associating with any SBS. This unique characteristic is a result of the logarithmic utility function of the simplified NBG model, where a user gains residual utility without associating. This phenomenon can explicitly motivate users to connect with an MBS since the additional gain in utility translates into a higher payoff for associating with an MBS. Therefore a modification of the objective function (5.2a) is proposed in the following with a penalty term:*

$$\sum_{m \in \mathcal{M}} \left[\log t_\tau - \log(1 - \frac{a_m}{w_m}) \right] - \sum_{m \in \mathcal{M}} P_m (1 - x_m), \tag{5.16}$$

where P_m is the penalty term, which a user pays when it does not choose to associate with an SBS. Setting $P_m = log(t_\tau)$ will cancel out the residual utility gain. Furthermore P_m can also be used to punish selfish nodes who do not want to cooperate. Thus it can be seen as a mechanism to induce cooperation among users, so that more users can be pushed towards the SBSs for content-retrieval purposes, given that the requested contents are available at the small-cell tier.

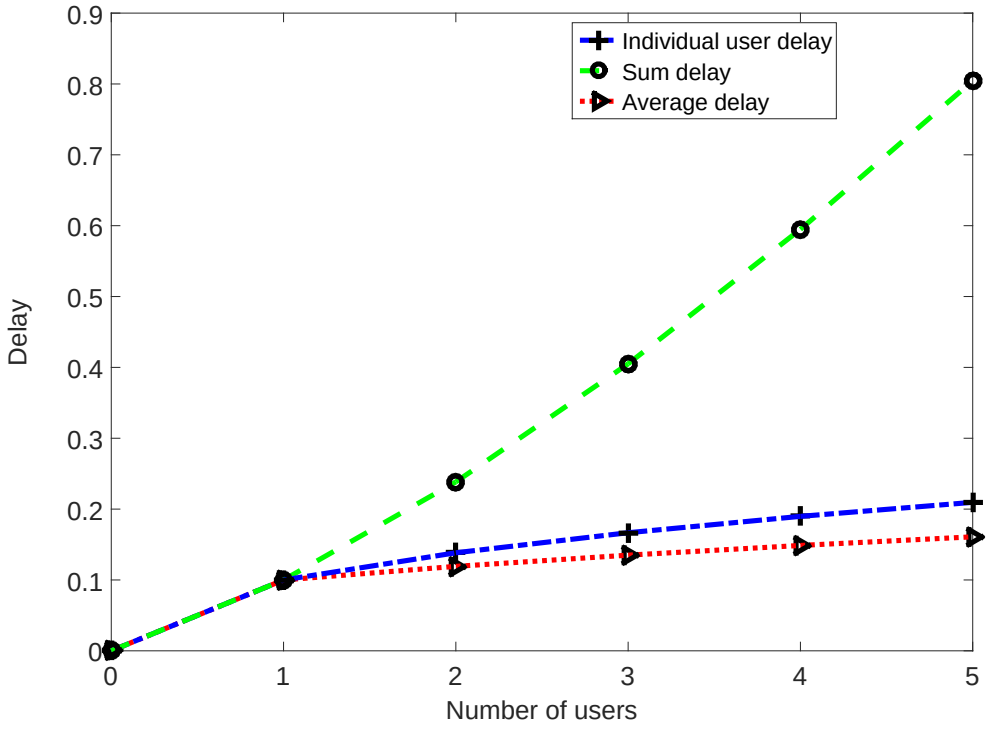


Figure 5.5: Delay performance of the HFC algorithm.

5.4 Numerical Analysis

This section analyses the performance of the proposed heuristic algorithm. The bandwidth allocation for users along with their delay performance is obtained by using MATLAB. The performance of the HFC algorithm is compared with the distributed solution proposed in Chapter 4. Before presenting the results, the values of the parameters used in simulations are listed: number of users 5, $t_\tau = 10$, $\Upsilon = 10$, number of contents $= 5$, $C_1 = 10$, $C_2 = 20$, $C_3 = 30$, $C_4 = 40$, $C_5 = 50$, $W^{max} = 100$ and $P_m = \log(t_\tau)$.

5.4.1 Delay Performance

We investigate the delay performance of the proposed HFC algorithm for given SBS bandwidth and threshold delay requirements. Using Algorithm 3, the SBS association

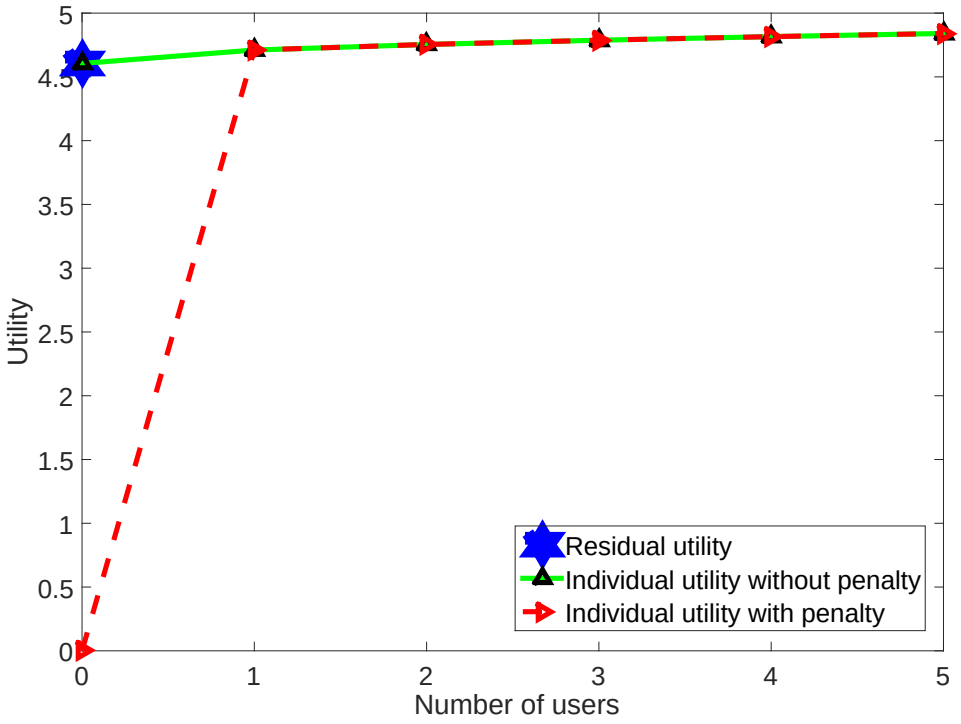


Figure 5.6: Utility Performance of HFC Algorithm.

and bandwidth allocation are computed for each user. The HFC algorithm allocates bandwidth to the five users as 10, 14.4536, 18, 21.0788 and 23.86. Once the bandwidth value is obtained, the delay for each user is calculated and plotted against the number of users as given by the red line in Fig. 5.5. The sum delay in SBS 1 and the average user delay per user is also calculated and plotted against the number of users. Both the sum delay and the average delay per user increases with the number of users.

5.4.2 Utility Performance

This section analyses the utility performance of the HFC algorithm as shown in Fig. 5.6. Using the bandwidth allocation obtained previously, we can calculate the utility for each user and plot it against the number of users. The green line plots the utility function

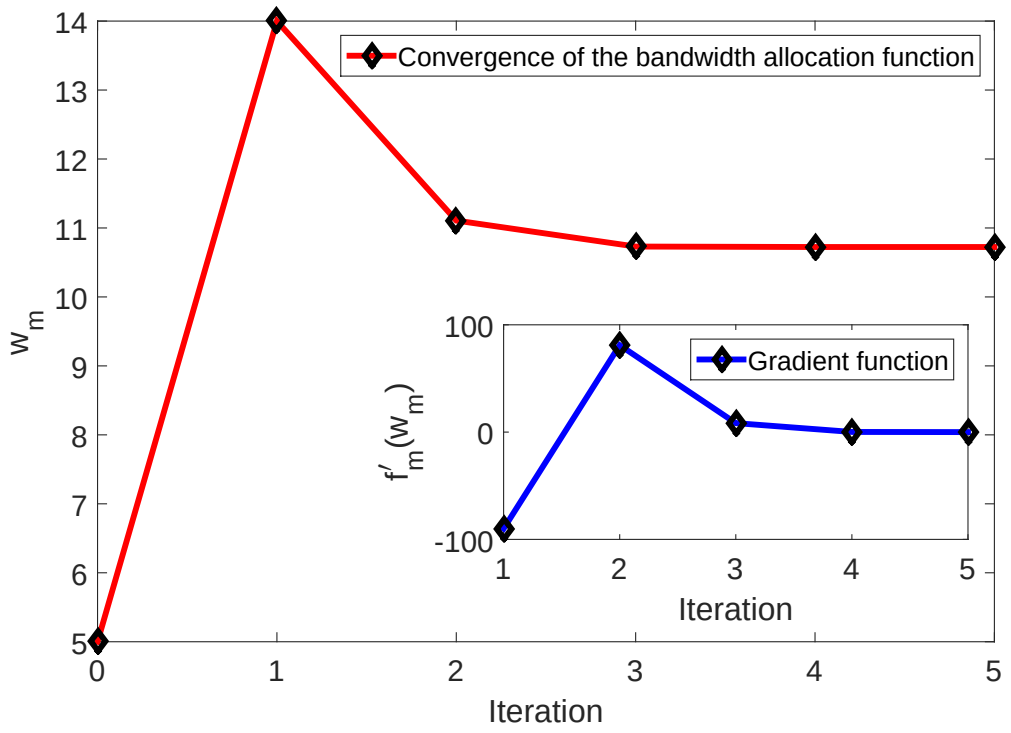


Figure 5.7: Convergence of the iterative bandwidth allocation.

(5.2a) and the red line plots the utility modification proposed in (5.16). The residual utility is given by the blue star in Fig. 5.6. It is clear that (5.2a) always assigns some utility value to a user even when a user is not associating with an SBS. Now consider that t_τ is much bigger than 10. In that case the residual utility can be significant since associating with an MBS will give a higher utility adding with the residual utility. As a result, including the penalty term in the utility function can exactly subtract the amount given by $\log(t_\tau)$

5.4.3 Convergence Analysis

We analyse the convergence properties of the iterative bandwidth allocation procedure given by Algorithm 2. The W^{max} value is 100 and $\epsilon = 0.001$. As observed in Fig. 5.7, the proposed algorithm only takes 5 iterations to converge to a feasible root. Thus

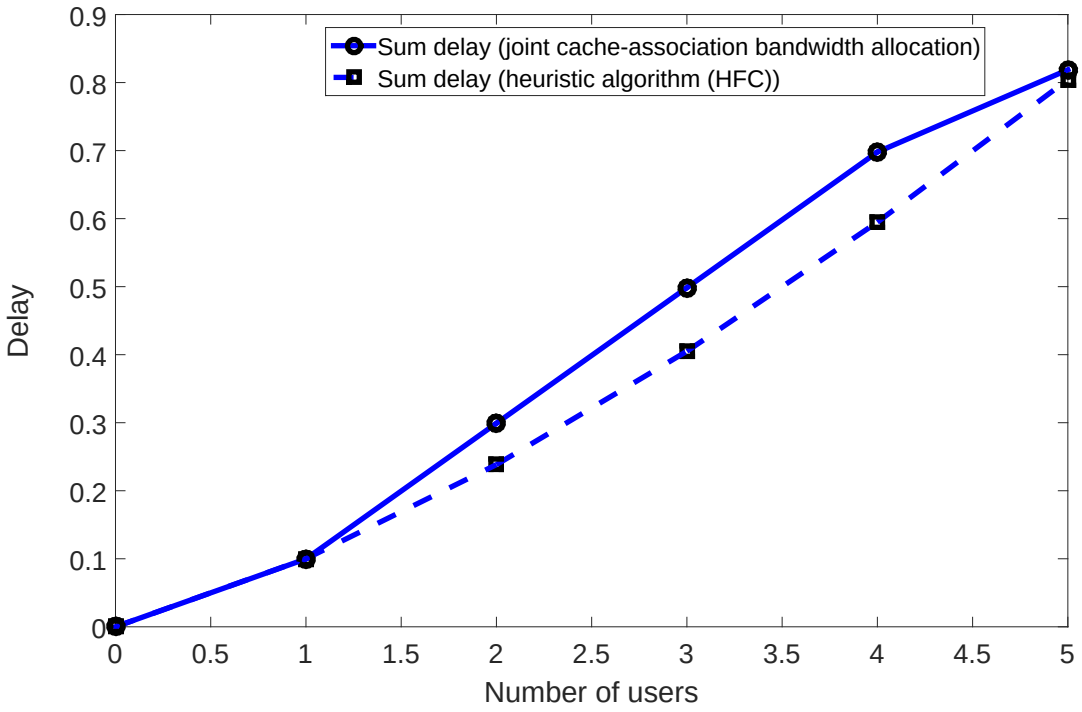


Figure 5.8: Sum delay performance.

convergence of the bandwidth allocation, i.e. of the HFC algorithm, is guaranteed.

5.4.4 Performance Comparison (Joint Optimisation Versus HFC)

This section compares the performance of the NBG-based joint cache-association and bandwidth-allocation algorithm proposed in Chapter 4 and the HFC algorithm, in terms of the delay performance of users. The simulation parameters are given in Section 5.4. Fig. 5.8 compares the sum-delay performance for both algorithms. It is clear that the HFC algorithm provides the better performance. The HFC algorithm's delay performance is superior until the 4-th user associates with an SBS, when the joint cache-association and bandwidth-allocation algorithm gives a sum delay of 0.7 and the HFC algorithm gives 0.6. When the 5-th user associates with the SBS, the difference between the sum-delay performance of both algorithms reduces. However the HFC algorithm still gives superior

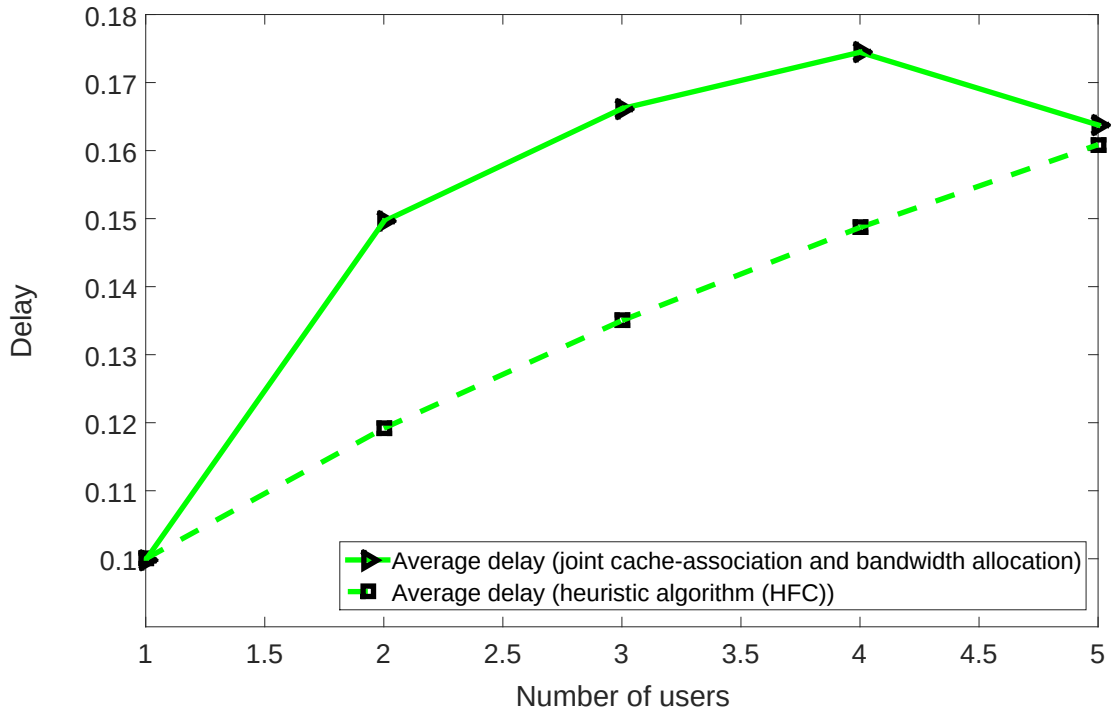


Figure 5.9: Average delay performance.

performance. Fig. 5.9 shows the average delay performance for both algorithms. The HFC algorithm clearly outperforms the joint cache-association and bandwidth-allocation algorithm, though the average delay performance of both algorithms reduces when the 5-th user joins the SBS. Note that the joint cache-association and bandwidth-allocation algorithm allocates only the minimum bandwidth to satisfy the delay requirement. On the other hand, the HFC algorithm's bandwidth allocation follows equation (5.10). As a result a performance gap exists until the fourth user association. Since the maximum bandwidth W^{max} is kept the same for both algorithm, the gap reduces significantly when the fifth user associates with an SBS. This is due to the fact that both algorithms allocate almost all of W^{max}.

5.5 Summary

This chapter provides a heuristic solution of the caching problem proposed in Chapter 4 where the joint optimisation of user association and the bandwidth allocation is decoupled into two separate problems. Thus the HFC algorithm executes in two stages where it assigns users to an SBS first; after that it allocates bandwidth to users. Interestingly the iterative bandwidth-allocation algorithm derived in this chapter follows Newton's method. Therefore a faster convergence is guaranteed than with the distributed solution. Numerical analysis shows that the HFC algorithm also outperforms the NBG-based joint cache-association and bandwidth-allocation algorithm in terms of delay performance.

Chapter 6

Fair Cache Association and Content Placement in 5G HetNet

6.1 Chapter Introduction

This chapter presents a content-caching problem which jointly considers the cache association and the cache placement in a 5G HetNet. We extend the caching model proposed in Chapter 4 and Chapter 5 by introducing the cache-placement decision in the objective-function formulation. This model replicates a scenario where a user's cache association depends on whether a content has been placed in the cache or not. In addition the threshold-delay constraint also needs to be satisfied for each user case. Both cache placement and cache association are modelled using the binary decision variables. The concept of NBG is applied to formulate an optimisation problem which results in a fair and Pareto-optimal solution. However the presence of both decision variables in the objective function and a constraint function complicates the optimisation problem. The new problem becomes a bilinear non-convex integer programming problem where the presence of bilinear term makes the problem hard to solve.

To solve the problem, the first step is to remove the mutual coupling between the binary decision variables. We attempt to introduce the concept of fairness into the content-caching problem, and as a result NBG-based optimisation is used. Note that FLPs are generally formulated into a traditional MINLP and they are analogous to the cache-association problem. In addition, FLPs do not consider fairness. Unlike traditional FLPs we incorporate the concept of fairness in our joint cache-placement and cache-association model. A linear relaxation method known as the McCormick relaxation [141], [142], [143] is applied, which allows us to search for a solution in the convex region. The bilinear term vanishes from the objective function as well as from the constraint after applying the relaxation. Thus, after the McCormick relaxation, the new problem becomes a convex problem [141], however it provides a near optimal solution. For mixed-integer nonlinear programming (MINLP) the relaxed problem may not provide an optimal solution. Such a case is demonstrated in [144], where the relaxed problem of the bargaining problem provides a near-optimal solution. In order to find a solution in a distributed manner, Lagrange relaxation is applied. It allows us to express the convexified problem as three subproblems. Thus each user needs to solve their own individual set of subproblems. Finally a subgradient method is applied to solve each subproblem iteratively. Numerical results are provided to demonstrate the performance of the proposed scheme. Due to the presence of bilinear terms as optimisation variables, finding the optimal solution of such MINLP is computationally exhaustive when the number of SBSs and users increases. Therefore, scalability of proposed bilinear programming is doubtful when the number of SBSs and users increases [145], [146]. To improve scalability, bilinear MINLP problems are solved using Lagrange relaxation and subgradient methods are generally used to avoid centralised computation [146]. Thus the distributed solution obtained by using Lagrange relaxation is solvable in polynomial or pseudo-polynomial time and therefore maintains tractability [126]. Though, subgradient method is applied to facilitate parallel computa-

Table 6.1: List of Important Symbol Notations

Symbol	Description
$\mathcal{I}$	Set of SBSs
$\mathcal{M}$	Set of users
$\boldsymbol{x}$	Vector of cache association variables
$\boldsymbol{y}$	Vector of cache-placement variables
$\mathcal{Q}$	Set of all contents cached by SBS
C_q	Size of the q-th content
w_{mi}	Allocated bandwidth to m-th user by i-th SBS
Υ_i	SNR of the i-th SBS
t_τ	Delay threshold
t_{miq}	Delay for m-th user
Θ_i	Cache size of i-th SBS
W^{min}	Minimum bandwidth requirement of each user
W_i^{max}	total bandwidth of SBS i
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$	Set of non-negative Lagrange multipliers

tion, the overall computation remains the same. Therefore to propose a computationally efficient algorithm a heuristic solution is needed which remains a future work of this chapter.

6.2 Network Model and Assumption

6.2.1 Notation Summary

The notation used is summarised in Table 6.1.

6.2.2 Elements of Wireless Networks

A HetNet is considered which consists of an MNO, where it deploys one MBS, and a large number of SBSs. Both the MBS and SBSs adopt LTE in their physical layer to provide services to users. The set of all SBSs is denoted by $\mathcal{I} = \{1, \cdots, I\}$ where $i \in \mathcal{I}$, denotes the i-th SBS. We denote the set of users as $\mathcal{M} = \{1, \cdots, M\}$ where $m \in \mathcal{M}$, denotes the m-th user. It is assumed that the MNO deploys SBSs very close to each other so that SBSs can serve as many content requests as possible. The MBS and SBSs are connected with the core network with high-capacity, and limited-capacity wired backhaul respectively.

Since spectrum is a scarce resource, spectrum sharing is preferred by the MNO. Therefore the MBS and SBSs share the same spectrum, i.e. they are deployed in an underlay spectrum-access manner. In order to avoid intra-tier and inter-tier interference, an interference coordination mechanism is considered among MBS and SBSs, in the downlink channels [114], [56], [115]. Therefore it is assumed that all SBSs deployed by the MNO provide an identical SNR. This assumption can be removed by devising a joint caching and transmission policy for delivering a content. We denote the SNR of the i-th SBS as Υ_i. Furthermore, to associate with an SBS, consider that the m-th user can negotiate with other users at the same time. These users are trying to associate with the nearby SBSs of the m-th user. Though a user can bargain with other users who are trying to associate with different SBSs, each user can only associate with one SBS. Thus the m-th user can only search for association with any SBS $i : i \in \mathcal{N}_m$, where $\mathcal{N}_m \subset \mathcal{I}$.

6.3 Problem Formulation

A global content library is defined by the set $\mathcal{Q} = \{1, \cdots, Q\}$. $|\mathcal{Q}|$ is the cardinality of set $\mathcal{Q}$ that defines the total content that can be cached by an SBS. However each SBS is able to place a subset of contents from $\mathcal{Q}$ restricted by their cache size, i.e. their storage

capacity. An MNO regulates contents and decide which contents are going to be included in a global pool. Let C_q denotes the size of the q-th content in bytes, $q \in \mathcal{Q}$. To calculate the cache capacity of an SBS we define a cache-association and cache-placement vector respectively as

$$\boldsymbol{x} = (x_{miq} \colon m \in \mathcal{M}, i \in \mathcal{N}_m, q \in \mathcal{Q}), \tag{6.1}$$

and

$$\boldsymbol{y} = (y_{iq} \colon i \in \mathcal{N}_m, q \in \mathcal{Q}), \tag{6.2}$$

where x_{miq} is the association decision of m-th user to the i-th SBS for the q-th content, y_{iq} is the content-placement decision of the i-th SBS for the q-th content. $x_{miq} \in \{0, 1\}$ is a binary decision variable which defines whether the m-th user requesting the q-th content is going to associate with the i-th SBS or not. $y_{iq} \in \{0, 1\}$ is a binary decision variable which defines whether the i-th SBS is going to place the q-th content or not. Each SBS's cache size is bounded by the following equation:

$$\sum_{q \in \mathcal{Q}} y_{iq} C_q \leq \Theta_i, \tag{6.3}$$

where Θ_i denotes the storage capacity of SBS i. Suppose that SBS i allocates a bandwidth denoted by w_{mi} for user m. Therefore the data rate of user m for content download purpose is given by the following equation:

$$R_{miq} = w_{mi} \log(1 + \Upsilon_i), \tag{6.4}$$

Suppose that the m-th user wants to retrieve the q-th content from the i-th SBS. The delay for retrieving the m-th content is:

$$t_{miq} = \frac{x_{miq} y_{iq} C_q}{w_{mi} \log(1 + \Upsilon_i)}, \tag{6.5}$$

Notice that the delay is dependent on the content size C_q and the allocated bandwidth w_{mi} for the m-th user at the i-th SBS. To simplify the delay equation given by (6.6), it is assumed that all users will be allocated the same amount of bandwidth, which is the minimum bandwidth to satisfy the threshold delay requirement. Assume that W^{min} is the minimum required bandwidth for all users. Substituting W^{min} in (6.6) we get:

$$t_{miq} = \frac{x_{miq} y_{iq} C_q}{W^{min} \log(1 + \Upsilon_i)}. \tag{6.6}$$

To associate with SBS i, the following delay requirement must be satisfied:

$$(t_\tau - t_{miq}) > 0, \tag{6.7}$$

where t_τ is the threshold delay, which is computed in terms of the average download delay per user, and t_{miq} is the delay for the m-th user for retrieving the q-th content from the i-th SBS. Therefore the m-th user will only associate with the i-th SBS when the content retrieval delay satisfies (6.7). The NBG-based optimisation for the proposed caching problem is given in Section 6.5 using (6.7) as a utility function.

6.4 Preliminaries on McCormick Relaxation

In order to apply linear relaxation for optimisation problems regarding bilinear products, McCormick envelopes [141], [143] are commonly used and the relaxation process is termed as McCormick relaxation. Consider variables $a, b \in \mathbb{R}$ whose bilinear product is ab. Each variable has a lower bound and an upper bound defined in the following form respectively:

$$a_L \leq a \leq a_U, \tag{6.8}$$

and

$$b_L \leq b \leq b_U. \tag{6.9}$$

Then the McCormick relaxation can be achieved by substituting the bilinear product ab with its surrogate $z : z \in \mathbb{R}$ in the optimisation problem. Furthermore the following envelopes are added as constraints

$$\begin{aligned} z \geq & a_L b + a b_L - a_L b_L, \\ z \geq & a_U b + a b_U - a_U b_U, \\ z \leq & a_U b + a b_L - a_U b_L, \\ z \leq & a b_U + a_L b - a_L b_U. \end{aligned} \tag{6.10}$$

6.5 Joint Cache-Association and Cache-Placement Algorithm

In this section, we model the cache-selection procedure into a Nash bargaining Game. In order to formulate the optimisation model in the form of a GNBS, we define the agreement utility function for user m as $U_m(x_{miq}, y_{iq}) := (t_\tau - \sum_{i \in \mathcal{N}_m} \sum_{q \in \mathcal{Q}} t_{miq})$. This implies that the m-th player wants to associate with any of its neighbouring SBS $i \in \mathcal{N}_m$ for retrieving the q-th content if and only if the content-retrieval delay remains under some threshold t_τ. Though each user can only associate with an SBS, it can simultaneously bargain with other users, who are competing for association with different SBSs. Note that if a user selects an SBS $i \in \mathcal{N}_m$ then it will not select any SBS $j \in \mathcal{N}_m$. Thus the association process satisfies the independence of irrelevant alternative axiom of the NBG. The disagreement utility for the m-th user is $U_m^0 := t_{min} = 0$. The interpretation of the disagreement utility is that the m-th user will not associate with the i-th SBS if the delay condition is not satisfied; i.e. if the download delay is higher or equal to the delay for MBS association, the deal between any m-th user and any i-th SBS will break down. Now

the joint optimisation problem is formulated as follows:

$$\max_{\boldsymbol{x},\boldsymbol{y}} \quad \sum_{m\in\mathcal{M}} \log\Big(U_m(x_{miq}, y_{iq}) - U_m^0\Big), \tag{6.11a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} y_{iq}C_q \le \Theta_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, \tag{6.11b}$$

$$\sum_{i\in\mathcal{N}_m} x_{miq}y_{iq} = 1, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.11c}$$

$$\sum_{q\in\mathcal{Q}} x_{miq} \le \frac{W_i^{max}}{W^{min}}, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.11d}$$

$$\sum_{i\in\mathcal{N}_m} t_{miq} \le t_\tau, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.11e}$$

$$x_{miq} \in \{0,1\}, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.11f}$$

$$y_{iq} \in \{0,1\}, q \in \mathcal{Q}, i \in \mathcal{N}_m. \tag{6.11g}$$

Notice that the ratio $\frac{W_i^{max}}{W^{min}}$ gives us the number of users who can be served by SBS i. Thus denote $S_i = \frac{W_i^{max}}{W^{min}}$, which is the number of users that SBS i serves. We substitute the value of the utility functions and obtain the following optimisation problem:

$$\max_{\boldsymbol{x},\boldsymbol{y}} \quad \log\Big(t_\tau - \sum_{i\in\mathcal{N}_m}\sum_{q\in\mathcal{Q}} t_{miq}\Big) \tag{6.12a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} y_{iq}C_q \le \Theta_i, q \in \mathcal{Q}, m \in \mathcal{M}, \tag{6.12b}$$

$$\sum_{i\in\mathcal{N}_m} x_{miq}y_{iq} = 1, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.12c}$$

$$\sum_{q\in\mathcal{Q}} x_{miq} \le S_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.12d}$$

$$\sum_{i\in\mathcal{N}_m} t_{miq} \le t_\tau, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.12e}$$

$$x_{miq} \in \{0,1\}, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.12f}$$

$$y_{iq} \in \{0,1\}, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{6.12g}$$

The optimisation problem presented by (6.12a)-(6.12g) is the proposed caching model formulated based on NBG, where each user's utility is generated by successfully downloading the desired content within the threshold delay. However each users' utility also depends on the decision of an SBS to place the content in its cache. Constraint (6.12b) imposes a restriction on maximum number of contents which can be cached at any SBS i must not exceed its cache capacity. Constraint (6.12c) explains that any user can only download a content from one SBS, given that SBS i decides to place the desired content of user m. Constraint (6.12d) imposes a restriction on a maximum association of SBS i, i.e. how many users can associate with a particular SBS depends on the total bandwidth of that SBS. Constraint (6.12e) states that the delay experienced by each must not exceed the threshold delay. Constraints (6.12f) and (6.12g) are binary decision variables for cache association and cache placement respectively.

The optimisation problem given by (6.12a)-(6.12f) involves two binary decision-making parameters. There exists a bilinear product in the objective function as well as in a constraint function, which makes the problem extremely hard to solve. However due to the characteristics of the GNBS, i.e. since it deals with the maximisation of logarithmic utility functions, it can be shown that under certain conditions the utility functions in GNBS can preserve concavity. Nevertheless the presence of bilinear terms in the objective function (6.12a), as well as in the constraint function (6.12c), makes both functions non-convex and hence the problem (6.12a)-(6.12f) is a non-convex integer programming problem. To comply with Nash's axioms, the problem must be transformed into convex programming. Therefore we apply a well-known linear relaxation technique involving bilinear terms called McCormick envelopes. Section 6.4 provides preliminary notes on McCormick relaxation. We apply these procedures to convert our problem into a convex programming in the following section.

6.5.1 Convex Relaxation Applying McCormick Relaxation

To obtain linear relaxations for bilinear terms $x_{miq}y_{iq}$, we introduce its surrogate z_{miq}. Applying McCormick relaxation, we obtain:

$$z_{miq} \geq 0, \tag{6.13}$$

$$z_{miq} \leq x_{miq}, \tag{6.14}$$

$$z_{miq} \leq y_{iq}, \tag{6.15}$$

and

$$z_{miq} \geq x_{miq} + y_{iq} - 1. \tag{6.16}$$

Adding (6.13)- (6.16) as constraints and replacing the bilinear product with its surro-

gate, we obtain the equivalent convex program as follows:

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad \sum_{m\in\mathcal{M}} \log\left[t_\tau - \sum_{i\in\mathcal{N}_m}\sum_{q\in\mathcal{Q}} \frac{z_{miq}C_q}{w_{mi}\log(1+\Upsilon_i)}\right], \tag{6.17a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} y_{iq}C_q \le \Theta_i, i\in\mathcal{N}_m, q\in\mathcal{Q}, \tag{6.17b}$$

$$\sum_{i\in\mathcal{N}_m} z_{miq} = 1, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17c}$$

$$\sum_{q\in\mathcal{Q}} x_{miq} \le S_i, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17d}$$

$$\sum_{i\in\mathcal{N}_m} \frac{z_{miq}C_q}{W^{min}\log(1+\Upsilon_i)} \le t_\tau, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17e}$$

$$x_{miq}\in[0,1], q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17f}$$

$$y_{iq}\in[0,1], q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17g}$$

$$z_{miq}\ge 0, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17h}$$

$$z_{miq}\le x_{miq}, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17i}$$

$$z_{miq}\le y_{iq}, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}, \tag{6.17j}$$

$$z_{miq}\ge x_{miq}+y_{iq}-1, q\in\mathcal{Q}, i\in\mathcal{N}_m, m\in\mathcal{M}. \tag{6.17k}$$

Now the above optimisation problem is solvable by applying Lagrange relaxation method which is described in the following section.

6.5.2 Distributed Solution: Lagrangian Relaxation Approach

This section details the Lagrange relaxation procedure for obtaining a distributed solution for the convexified problem (6.17a)-(6.17k). We relax constraints (6.17i), (6.17j) and

(6.17k) and obtain the following Lagrange function:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \\ \sum_{m\in\mathcal{M}}\log\left[t_\tau - \sum_{i\in\mathcal{N}_m}\sum_{q\in\mathcal{Q}}\frac{z_{miq}C_q}{W^{min}\log(1+\Upsilon_i)}\right]+ \\ \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{N}_m}\sum_{q\in\mathcal{Q}}\Big[\alpha_{miq}(z_{miq}-x_{miq})+\beta_{miq}(z_{miq}-y_{iq})+ \\ \gamma_{miq}(x_{miq}+y_{iq}-z_{miq}-1)\Big],\end{aligned} \tag{6.18}$$

where $\boldsymbol{\alpha}\succeq 0, \boldsymbol{\beta}\succeq 0, \boldsymbol{\gamma}\succeq 0$ are the nonnegative KKT multipliers for the constraints. We rearrange the Lagrangian (6.18) in the following form:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \\ \sum_{m\in\mathcal{M}}\left[\log\left(t_\tau - \sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}\frac{z_{miq}C_q}{W^{min}\log(1+\Upsilon_i)}\right)+\sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}z_{miq}(\alpha_{miq}+\right. \\ \left.\beta_{miq}+\gamma_{miq})\right]+\sum_{m\in\mathcal{M}}\left[\sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}x_{miq}(-\alpha_{miq}+\gamma_{miq})\right]+ \\ \sum_{m\in\mathcal{M}}\left[\sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}(-y_{iq}\beta_{miq}+y_{iq}\gamma_{miq}-\gamma_{miq})\right] \\ \Rightarrow \mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \\ \sum_{m\in\mathcal{M}}\Big[F'(z_{miq})+G'(x_{miq})+H'(y_{iq})\Big],\end{aligned} \tag{6.19}$$

where,

$$F'(z_{miq}) = \left[\log\left(t_\tau - \sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}\frac{z_{miq}C_q}{W^{min}\log(1+\Upsilon_i)}\right)+\sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}z_{miq}(\alpha_{miq}+\beta_{miq}+\gamma_{miq})\right], \tag{6.20}$$

$$G'(x_{miq}) = \left[\sum_{i\in\mathcal{N}m}\sum_{q\in\mathcal{Q}}x_{miq}(-\alpha_{miq}+\gamma_{miq})\right], \tag{6.21}$$

and

$$H'(y_{iq}) = \Bigg[\sum_{i \in \mathcal{N}m} \sum_{q \in \mathcal{Q}} (-y_{iq}\beta_{miq} + y_{iq}\gamma_{miq} - \gamma_{miq}) \Bigg]. \tag{6.22}$$

We can write the dual function as

$$d(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}). \tag{6.23}$$

Now the dual problem can be written as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} \quad d(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \tag{6.24a}$$

$$\text{s.t.} \quad \sum_{q \in \mathcal{Q}} y_{iq} C_q \leq \Theta_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, \tag{6.24b}$$

$$\sum_{i \in \mathcal{N}_m} z_{miq} = 1, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.24c}$$

$$\sum_{q \in \mathcal{Q}} x_{miq} \leq S_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.24d}$$

$$\sum_{i \in \mathcal{N}_m} \frac{z_{miq} C_q}{W^{min} \log(1 + \Upsilon_i)} \leq t_\tau, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.24e}$$

$$x_{miq} \in [0, 1], q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.24f}$$

$$y_{iq} \in [0, 1], q \in \mathcal{Q}, i \in \mathcal{N}_m, \tag{6.24g}$$

$$z_{miq} \geq 0, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{6.24h}$$

Notice that the Lagrangian function (6.19) exhibits an especial structure. Utilising it, the dual problem is separable in terms of $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus (6.24a)-(6.24h) is decomposed into three subproblems.

Subproblem 1:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} \quad \sum_{m\in\mathcal{M}} F'(z_{miq}), \tag{6.25a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} z_{miq} = 1, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.25b}$$

$$\sum_{i\in\mathcal{N}_m} \frac{z_{miq} C_q}{W^{min}\log(1+\Upsilon_i)} \leq t_\tau, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.25c}$$

$$z_{miq} \geq 0, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{6.25d}$$

Subproblem 2:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \quad \sum_{m\in\mathcal{M}} G'(x_{miq}) \tag{6.26a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} x_{miq} \leq S_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}, \tag{6.26b}$$

$$x_{miq} \in [0,1], q \in \mathcal{Q}, i \in \mathcal{N}_m, m \in \mathcal{M}. \tag{6.26c}$$

Subproblem 3:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad \sum_{m\in\mathcal{M}} H'(y_{iq}), \tag{6.27a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} y_{iq} C_q \leq \Theta_i, q \in \mathcal{Q}, i \in \mathcal{N}_m, \tag{6.27b}$$

$$y_{iq} \in [0,1], q \in \mathcal{Q}, i \in \mathcal{N}_m. \tag{6.27c}$$

Notice that after applying McCormick relaxation the bilinear term vanishes. Performing dual-decomposition further reduces the complexity of solving a large optimisation problem by decomposing the convexified problem into separate subproblems. Instead of solving three subproblems centrally we further separate the subproblems for each user. Thus any m-th user has to solve the following subproblems:

Subproblem 1 of m-th User:

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}{min} \quad F'(z_{miq}), \tag{6.28a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}_m} z_{miq} = 1, \tag{6.28b}$$

$$\sum_{i\in(N_m} \frac{z_{miq}C_q}{W^{min}\log(1+\Upsilon_i)} \leq t_\tau, \tag{6.28c}$$

$$z_{miq} \geq 0. \tag{6.28d}$$

Subproblem 2 of m-th User:

$$\underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{min} \quad G'(x_{miq}), \tag{6.29a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} x_{miq} \leq S_i, \tag{6.29b}$$

$$x_{miq} \in [0,1]. \tag{6.29c}$$

Subproblem 3 of m-th User:

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{min} \quad H'(y_{iq}), \tag{6.30a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}} y_{iq}C_q \leq \Theta_i, \tag{6.30b}$$

$$y_{iq} \in [0,1]. \tag{6.30c}$$

Thus Lagrangian relaxation provides a powerful tool to decompose a centralised optimisation problem into independent user problems. Here we apply the subgradient method to compute and update the dual variables $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In the k-th iteration the dual variables are updated as follows:

$$\alpha_{miq}(k+1) = [\alpha_{miq}(k) + \rho_\alpha(k)g_\alpha(k)]^+, \tag{6.31}$$

where $\alpha_{miq}(k)$ is the k-th iterate, $g_\alpha(k)$ is any subgradient at $\alpha_{miq}(k)$ and $\rho_\alpha(k)$ is an appropriate step size of the k-th iteration. The step-size selection is detailed in Section 4.4.2 of Chapter 4.

$$\beta_{miq}(k+1) = [\beta_{miq}(k) + \rho_\beta(k) g_\beta(k)]^+, \tag{6.32}$$

where $\beta_{miq}(k)$ is the k-th iterate, $g_\beta(k)$ is any subgradient at $\beta_{miq}(k)$ and $\rho_\beta(k)$ is an appropriate step size of the k-th iteration.

$$\gamma_{miq}(k+1) = [\gamma_{miq}(k) + \rho_\gamma(k) g_\gamma(k)]^+, \tag{6.33}$$

where $\gamma_{miq}(k)$ is the k-th iterate, $g_\gamma(k)$ is any subgradient at $\gamma_{miq}(k)$ and $\rho_\gamma(k)$ is an appropriate step size of the k-th iteration. The processing delay of solving the independent user problem is the same as solving a Lagrangian relaxation, which can be solved in polynomial time. Therefore each of the subproblems is solvable in polynomial time.

6.6 Numerical Results

This section analyses the numerical results of the joint cache placement and association problem in terms of user association and average delay performance. The caching performance is numerically simulated for a caching-enabled HetNet, which is composed of two SBSs and one MBS. The content library has two contents and each SBS can only place one of the contents according to their individual cache capacity. It is assumed that the content request made by each user to both SBSs follows Zipf's distribution. The role of content distribution is very important to maximise the cache hit rate at an SBS. To compute and compare SBS cache hit rate, Zipf distribution is necessary since it models a system when there are unobserved and underlying variables. Zipf distribution is a most widely used model [72], [73], [122] for calculating cache hit rate which arises from Zipf's law, describing a relation between rank order and frequency of occurrence [147]. It states

that, when observations are ranked by their frequency, the frequency of a particular observation is inversely proportional to its rank [147]. Thus Zipf distribution describes the popularity of a content where a content is ranked first and then its frequency, i.e. how many times a content is requested can be can calculated by Zipf's law. Zipf's distribution states that the relative probability for requesting the q-th content request is proportional to $\frac{1}{q^\eta}$ where η is a positive value and $0 \leq \eta \leq 1$. Thus the Zipf's distribution can explicitly determine the content popularity. The Zipf's distribution function is expressed as follows [148]:

$$r_q = \frac{1}{q^\eta}\left(\sum_{q=1}^{|\mathcal{Q}|}\frac{1}{q^\eta}\right)^{-1}, q = 1, 2, \tag{6.34}$$

where r_q denotes the request generated for the q-th content and $|\mathcal{Q}|= 2$ for our scenario. Here $q = 1$ means that content 1 is more popular than content 2 and will be requested more frequently. The content sizes are $C_1 = 115, C_2 = 100$. The minimum bandwidth for each association is $W^{min} = 20$ and the bandwidth of each SBS is 100. The threshold delay $t_\tau = 10$.

Now we generate content-download requests for 20 users and observe the user association to each SBS for content 1 and content 2. It is important to examine the potency of the proposed utility function in terms of SBS association, since it is the goal of this chapter to increase SBS association for content caching. Therefore we will investigate the user association in each SBS for different user settings and given system parameters. We use five different user settings to analyse the performance of the content-caching system. The five different user settings are: 1-user, 5-user, 10-user, 15-user and 20-user. The NBG-based caching algorithm will give bandwidth allocation and optimal association, and plotting against various user settings will allow us to calculate the user association in %. This gives us a user bound, i.e. the number of users who can be served by an SBS, depending on the cache availability, for given system parameters.

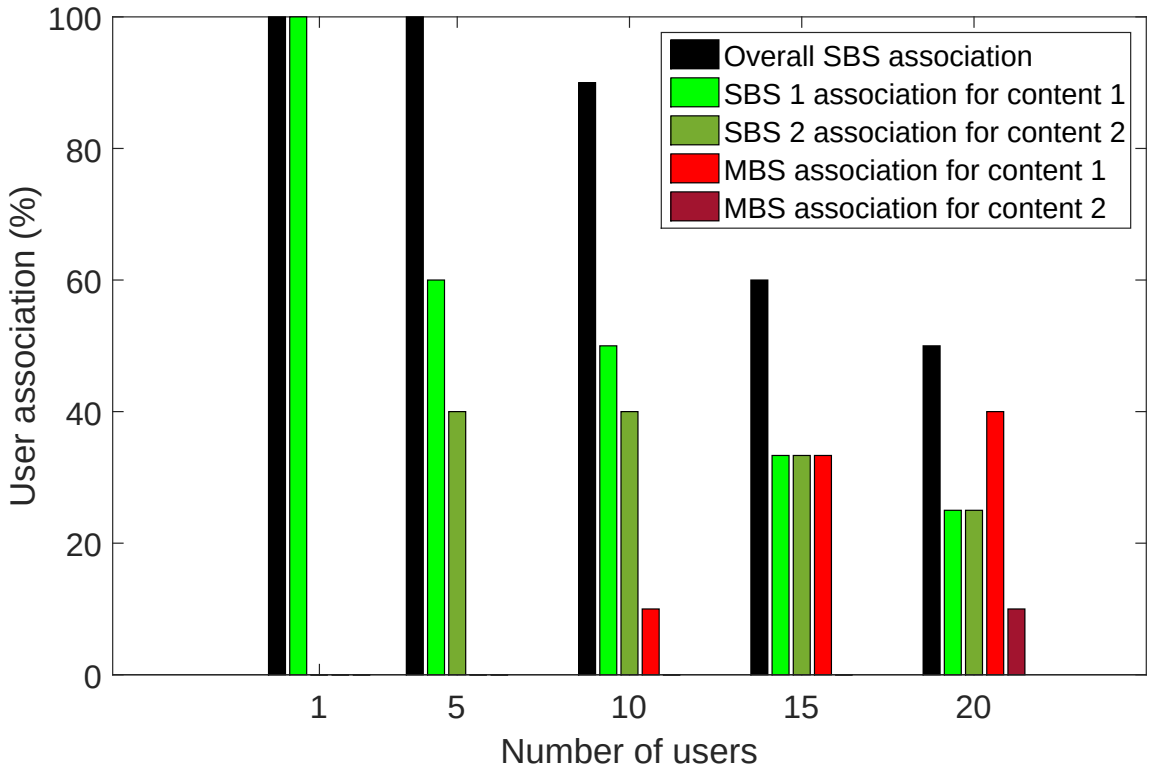


Figure 6.1: User association considering cache placement when $W^{min} = 20$.

Thus for a given number of users we calculate user association for each content to each BS. The result obtained is plotted in Fig. 6.1. Notice that content 1 is requested more than content 2. This is due to the fact that content 1 ranks higher than content 2 according to Zipf's distribution. Notice that in 10-user setting, one user requesting content 1 joins MBS. This phenomenon occurs since SBS 1 can not accommodate more users without violating the minimum bandwidth constraint. For SBS 2, this situation occur in 20-user setting since fewer number of users request content 2.

This gives us an idea to design a content caching system using content popularity. Suppose that an SBS can place more than 1 content. Placing multiple popular contents will certainly increase SBS association if W^{min} is small. Such a case is demonstrated in Fig. 6.2, where $W^{min} = 10$. Notice that SBS 2 is able to serve all the content request.

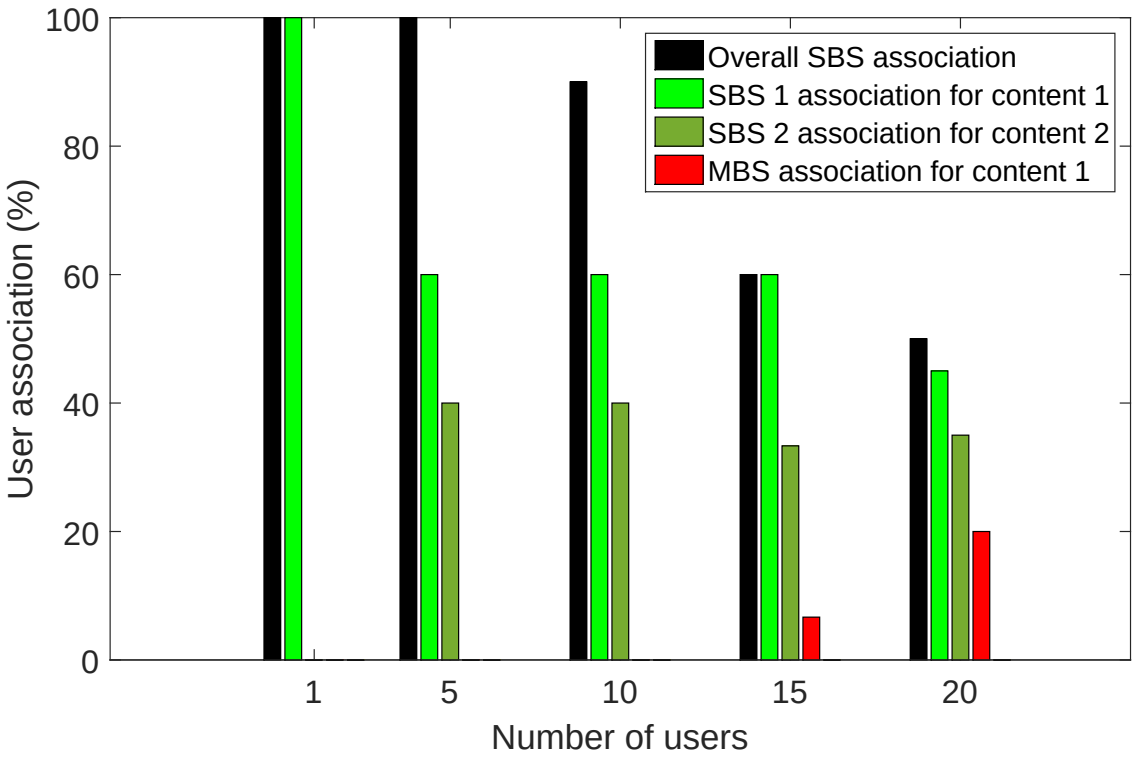


Figure 6.2: User association considering cache placement when $W^{min} = 10$.

This will increase individual user delay but more users can be assigned to SBS as long as threshold delay constraint is satisfied.

Remark 3. *An important observation is that a decision of an SBS about placing the most popular content will increase its cache association. As a result it can overload an SBS in this process. On the other hand if there are multiple SBSs and if all of them decide to place the most popular content, then some SBSs are going to be under-utilised. Thus there exists a trade-off between content popularity and efficient SBS utilisation. Therefore the content-placement decision needs to be carefully optimised for efficient SBS utilisation. On the other hand, utilising t_τ congestion can be avoided and SBS utilisation can be improved.*

Fig. 6.3 shows the average data rate performance of users associated with SBS 1 and SBS 2. It has been found that users requesting content 1 suffers from higher average

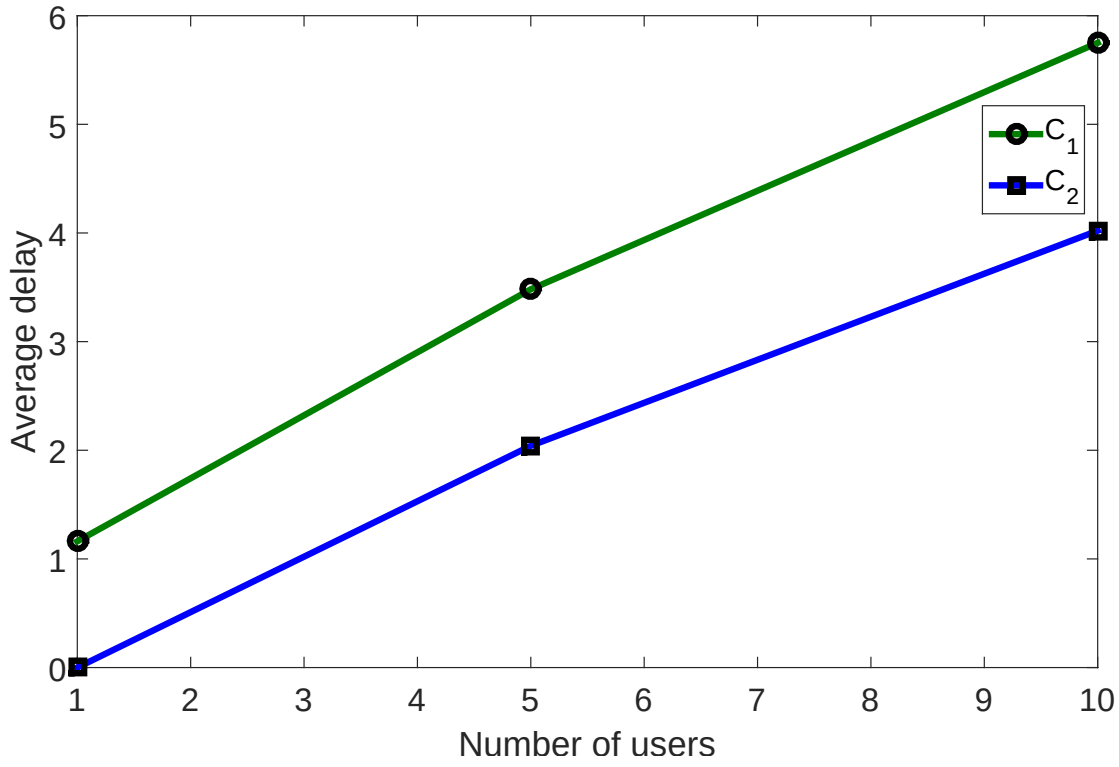


Figure 6.3: Average delay performance of users.

delay. Notice that at the 10-user setting SBS 1 serves 5 users and SBS 2 serves 4 users. In addition, content 1 has a bigger size than content 2. As a result, users requesting content 1 experience more delay.

6.7 Summary

In this chapter, an NBG-based joint content-placement and user-association problem is considered. It is found that the formulated optimisation problem is non-convex due to the presence of the bilinear terms. Therefore, a linear relaxation method called McCormick relaxation is first used to convexify the problem. After that Lagrange relaxation method is applied to decompose the centralised problem into independent user problems. Finally a small-scale optimisation of the proposed system is performed using MATLAB. The

findings are plotted for different user settings to show how the proposed system works for given system parameters. The idea, presented in the form of an NBG in this chapter, can be improved further if SBSs cooperate to maximise caching opportunity.

Chapter 7

Thesis Conclusion and Future Work

This chapter presents a summary of the research outcomes of this thesis. In addition, brief notes on the direction of future work are provided.

7.1 Thesis Conclusions

This thesis focuses on the key enabling technologies for 5G HetNets. As explained in the literature review, both offloading and content-caching technologies provide an inexpensive upgrade to enable exciting network features and enhance the data rate of users without relying on new spectrum. It is clearly pointed out in the recent literature that legacy resource-allocation schemes fail to fully utilise precious RF spectrum. As a consequence they give an inefficient spectrum utilisation in the small-cell tier. Thus offloading and the content-caching mechanisms fit nicely into the pool of prospective technologies to unlock the potential of small-cells. To engineer an inexpensive and efficient solution for the existing spectrum-scarcity problem, this thesis proposes novel resource allocation techniques for 5G HetNets, summarised as follows:

- A price-based offloading scheme is proposed for macrocell-controlled offloading in

Chapter 3, where each user performs network selection based on EGT. The pricing scheme proposed in Chapter 3 is viewed as a control mechanism from the macrocell perspective and has two objectives. Firstly it enables macrocell-controlled offloading, i.e. a macrocell controls its population share by means of pricing. A macrocell first adopts linear pricing, which reduces with an increasing number of macrocell users. As soon as the macrocell PS increases to a certain proportion, it triggers a high threshold price for users. It has been observed that the effect of a threshold price is more severe to users who suffer from interference from nearby SBSs. There arises a situation for a user when associating with the SBS, i.e. the interferer, becomes more beneficial in terms of the utility gain than remaining associated with a macrocell. An EGT-based network-selection algorithm is provided which shows that a significant proportion of users will select small-cells under the proposed pricing strategies. The existence, uniqueness and stability criteria of the proposed algorithm are derived. Various network configurations have been tested to analyse the flexibility, scalability and efficiency of the proposed pricing algorithm. Simulation results showed excellent performance in terms of offloading users efficiently to SBSs and it certainly outperforms the contemporary RSS-based user-association technique.

- Video streaming services and video-on-demand services have gained massive popularity among consumers in recent times. As a consequence video contents are adding up to 70% of overall mobile data traffic. As a result an MBS cannot guarantee high quality and sustainable performance to all users. Therefore a new outlook is required on content-based services, which should focus on SBS-based caching rather than MBS-based caching. Considering this fact in Chapter 4, a novel caching strategy is developed to maximise the locally available caching opportunities provided by small-cells. A novel utility function is designed to ensure SBS association for content downloading. The inclusion of an MBS's average content-download delay in

the utility-function design enables an MNO to devise a SBS-centric content-caching algorithm. An NBG-based algorithm is proposed to ensure fair allocation of resource to the associated users. Both centralised and distributed algorithms are developed for the proposed NBG-based optimisation scheme. Numerical results are provided outlining the trade-off between the threshold delay and the SBS bandwidth. Simulation results also show a clear benefit for utilising a threshold delay to maximise SBS association. In addition two different service strategies are proposed based on varying the threshold delay, where a flexible threshold delay can be implemented for services such as video-on-demand and a stringent threshold delay can be used for services such as video streaming. In summary the proposed algorithm provides a framework for SBSs to play a central role in content caching in a 5G HetNet.

- A major pitfall of the NBG-based joint optimisation scheme proposed for cache association and fair bandwidth allocation in Chapter 4 is that it comes with an inherent computational complexity. The distributed solution proposed using Lagrange relaxation is solvable in polynomial time. However it has a very high computational overhead. To solve that problem, a heuristic algorithm is proposed in Chapter 5 where the association and bandwidth allocation are solved separately. It is shown that the HFC algorithm allocates bandwidth to users in a waterfilling-like manner. It has been analytically proved that the convergence of the bandwidth allocation follows that of Newton's method. As a result, a faster convergence is guaranteed when using the distributed algorithm proposed in Chapter 4.

- The caching algorithm proposed in Chapter 6 considers a scenario where associating with an SBS for content downloading depends on the content-placement decision of SBSs. A joint optimisation problem is formulated into an NBG-based game-theoretic optimisation model. Due to the presence of bilinear terms in the objective

function as well as in the constraint function, the problem becomes a non-convex nonlinear integer programming problem. To find an optimal solution, each bilinear term is replaced by its convex envelopes found by applying McCormick relaxation. The convexified problem is then solved by using Lagrange relaxation. A distributed algorithm is proposed based on the subgradient method.

7.2 Future Research Directions

Both offloading and content-caching techniques are regarded as potential candidate technologies to be featured in 5G networks. This thesis has investigated those topics from an engineering mindset to provide low-cost, efficient and robust networking solutions. Offloading and content-caching technologies have the capabilities to transform the wireless network experience to the next level without expanding the spectrum domain (mm-wave) or applying complex technology such as massive-MIMO to harness the SE gain. Inspired by the HetNet potential and spectrum-sharing benefits, this thesis provides a novel price-based offloading scheme and a threshold-delay-based utility-function design for content caching. Both schemes can be used as a benchmark to further enhance the efficiency of proposed schemes so that they can be improved to develop state-of-the-art offloading and content-caching schemes respectively. In addiction, the application of game-theory-based analytical methodology can be used as a framework to instill the much-desired *network intelligence* in future wireless networks. Based on the research findings, several aspects of the algorithms can be improved further.

- The offloading scheme proposed in Chapter 3 achieves the design objectives. It has been shown that the macrocell has the much-needed *control factor* to maintain its optimum population share. However the threshold data rate and pricing parameters for small-cell association and macrocell association can be optimised to give an

optimum population share in each tier. Furthermore appropriate software and test-beds need to be designed for field trials.

- The content-caching schemes proposed in Chapters 4-6 can be utilised to develop various content-based services, where the threshold delay can be chosen based on the type of content request made by the users. In other words the efficiency of the edge caching can be enhanced greatly by an appropriate choice of threshold delay. Co-located SBSs can form clusters and jointly decide on placing different category of contents in a cooperative manner. Forming clusters to place and deliver contents is an excellent alternative to tackle the content diversity issue rather than placing a large number of SBSs. In this way, each SBS is able to place different types of contents rather than placing the same contents in all SBSs, via cooperation and coordination. This will allow users to download a large portion of contents from the small-cell tier, which will eventually increase the small-cell cache hit rate. Therefore an effective clustering or SBS grouping scheme is needed to deliver an optimal performance in SBS-based caching. Furthermore there exists a challenge to develop a faster algorithm so that real-time implementation is possible for an SBS-based content-retrieval system. A delay can arise in the form of congestion, queuing delay, backhaul delay. In addition, delay can also arise if a requested content is not available in an SBS. Then the SBS can either redirect the requesting user to a nearby SBS or the SBS can download the file itself and broadcast it to the user who requested the content. Therefore the delay model proposed in Chapter 4 can be upgraded to a more refined delay model. Thus the schemes proposed in Chapters 4-6 can be used a framework and they can be improved further for developing a full-fledged content-caching system for 5G networks.

Appendix A

List of Acronyms

AAA	Authorization and accounting
AI	Artificial intelligence
APs	Access points
BSs	Base stations
BRUTE	Best responding user association with traffic estimation
CAPEX	CAPital EXpenditure
CPs	Content providers
CRE	Cell range expansion
CR	Cognitive radio
CRN	Cognitive radio networks
EGT	Evolutionary game theory
EPC	Evolved pocket core
ESS	Evolutionary stable strategy
FLP	Facility location problem
1G	First generation
4G	Fourth generation

5G	Fifth generation
GAP	Generalised assignment problem
GNBS	Generalised Nash bargaining solution
HetNets	Heterogenous networks
ICN	Information centric networking
ILP	Integer linear programming problem
IoT	Internet of things
LTE	Long term evolution
MBS	Macrocell base station
MNOs	Mobile network operators
MU	Macro users
NBG	Nash bargaining game
NBS	Nash bargaining solution
NE	Nash equilibrium
ODE	Ordinary differential equations
OPEX	OPerating EXpenditure
PRBs	Physical resource blocks
QoE	Quality of experience
QoS	Quality of service
RF	Radio frequency
RSS	Received signal strength
SBSs	Small cell base stations
SE	Spectral efficiency
SO	Self organizing
SON	Self organized networks
SINR	Signal to interference plus noise ratio

SNR	Signal to noise ratio
SP	Service provider
2G	Second generation
3G	Third generation
UEs	User equipments
URLLC	Ultra reliable and low latency communications
VLC	Visible light communication
WISDOM	Wireless innovative system for dynamic operating megacommunications

References

[1] A. Bell, "Apparatus for signaling and communicating called photophone," *US Patent 235199*, vol. PP, no. 99, pp. 1–1, Dec. 1880.

[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021," Tech. Rep., February 2017, online:https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf, accessed 2018-02-11.

[3] R. Baldemair, E. Dahlman, G. Fodor, G. Mildh, S. Parkvall, Y. Selen, H. Tullberg, and K. Balachandran, "Evolving wireless communications: Addressing the challenges and expectations of the future," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 24–30, Mar. 2013.

[4] U. Varshney, "4g wireless networks," *IT Professional*, vol. 14, no. 5, pp. 34–39, Sept 2012.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[6] "LTE for utilities: supporting smart grids," Tech. Rep., 2013, online:https://vdocuments.com.br/download/white-paper-lte-for-utilities-supporting-smart-grids, accessed 2018-06-29.

[7] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.

[8] "Ericsson mobility report," Tech. Rep., June 2017, online:https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-june-2017.pdf, accessed 2018-06-29.

[9] P. Olanders, "Network evolution towards 5g," Tech. Rep., 2015, online:https://www.atis.org/wsts/2015/papers/1-0_Ericsson_Olanders_Network%20Evolution%20Towards%205G.pdf, accessed 2018-02-11.

[10] C. Badoi, N. Prasad, V. Croitoru, and R. Prasad, "5g based on cognitive radio," *Wireless Personal Communications*, vol. 57, no. 3, pp. 441–464, Apr. 2011.

[11] T. Janevski, "5g mobile phone concept," in *2009 6th IEEE Consumer Communications and Networking Conference*, Jan 2009, pp. 1–2.

[12] F. Hu, B. Chen, and K. Zhu, "Full spectrum sharing in cognitive radio networks toward 5g: A survey," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2018.

[13] C. X. Wang, F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5g wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, February 2014.

[14] M. Z. Chowdhury, M. T. Hossan, A. Islam, and Y. M. Jang, "A comparative survey of optical wireless technologies: Architectures and applications," *IEEE Access*, vol. 6, pp. 9819–9840, 2018.

[15] H. Haas and C. Chen, *Visible Light Communication in 5G*, 1st ed. United States: Cambridge University Press, 2016.

[16] N. Saha, M. S. Ifthekhar, N. T. Le, and Y. M. Jang, "Survey on optical camera communications: challenges and opportunities," *IET Optoelectronics*, vol. 9, no. 5, pp. 172–183, 2015.

[17] P. Rysavy, "Challenges and considerations in defining spectrum efficiency," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 386–392, Mar. 2014.

[18] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[19] Y. Liang, A. Goldsmith, G. Foschini, R. A. Valenzuela, and D. Chizhik, "Evolution of base stations in cellular networks: Denser deployment versus coordination," in *2008 IEEE International Conference on Communications*, May 2008, pp. 4128–4132.

[20] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5g networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second quarter 2016.

[21] C. Yang, J. Li, and M. Guizani, "Cooperation for spectral and energy efficiency in ultra-dense small cell networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 64–71, Feb. 2016.

[22] E. Hossain, L. B. Le, and D. Niyato, *Radio Resource Management in Multi-Tier Cellular Wireless Networks*. Hoboken, NJ, USA: Wiley, 2013.

[23] J. S. Lin and K. T. Feng, "Femtocell access strategies in heterogeneous networks using a game theoretical framework," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1208–1221, Mar. 2014.

[24] D. Xu, Y. Li, X. Chen, J. Li, P. Hui, S. Chen, and J. Crowcroft, "A survey of opportunistic offloading," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2198–2236, thirdquarter 2018.

[25] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5g-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

[26] S. Yunas, W. Ansari, and M. Valkama, "Technoeconomical analysis of macrocell and femtocell based hetnet under different deployment constraints," *Mobile Information Systems*, vol. 2016, 2016.

[27] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 336–361, First 2013.

[28] S. Kim and P. K. Varshney, "Adaptive load balancing with preemption for multimedia cellular networks," in *2003 IEEE Wireless Communications and Networking, 2003. WCNC 2003.*, vol. 3, March 2003, pp. 1680–1684 vol.3.

[29] J. S. Harsini and F. Lahouti, "Adaptive transmission policy design for delay-sensitive and bursty packet traffic over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 776–786, Feb 2009.

[30] N. Samaan and A. Karmouch, "Towards autonomic network management: an analysis of current and future research directions," *IEEE Communications Surveys Tutorials*, vol. 11, no. 3, pp. 22–36, rd 2009.

[31] J. Suzuki and T. Suda, "A middleware platform for a biologically inspired network architecture supporting autonomous and adaptive applications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 249–260, Feb 2005.

[32] P. Demestichas, A. Georgakopoulos, K. Tsagkaris, and S. Kotrotsos, "Intelligent 5g networks: Managing 5g wireless/mobile broadband," *IEEE Vehicular Technology Magazine*, vol. 10, no. 3, pp. 41–50, Sept 2015.

[33] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5g: When cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, October 2017.

[34] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5g: how to empower son with big data for enabling 5g," *IEEE Network*, vol. 28, no. 6, pp. 27–33, Nov 2014.

[35] X. Lin, J. Andrews, and A. Ghosh, "Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *Communications, IEEE Transactions on*, vol. 61, no. 9, pp. 4002–4015, Sep. 2013.

[36] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5g systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[37] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5g cloud radio access networks," *IEEE Network*, vol. 31, no. 4, pp. 35–41, July 2017.

[38] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks:issues, measures and challenges," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 5–24, First 2014.

[39] J. Sangiamwong, Y. Saito, N. Miki, T. Abe, S. Nagata, and Y. Okumura, "Investigation on cell selection methods associated with inter-cell interference coordination in heterogeneous networks for lte-advanced downlink," in *17th European Wireless 2011 - Sustainable Wireless Technologies*, Apr. 2011, pp. 1–6.

[40] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5g multi-tier cellular wireless networks: An interference management perspective," *Wireless Communications, IEEE*, vol. 21, no. 3, pp. 118–127, Jun. 2014.

[41] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Outage probability for heterogeneous cellular networks with biased cell association," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec. 2011, pp. 1–5.

[42] S. Brams, *Negotiation Games: Applying Game Theory to Bargaining and Arbitration*. Routledge, 2003.

[43] C. Courcoubetis and R. Weber, *Pricing Communication Networks: Economics, Technology and Modelling*. Chichester, UK: John Wiley & Sons Ltd, 2003.

[44] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[45] "Minimum requirements related to technical performance for imt2020 radio interface(s)," Tech. Rep., Nov. 2017, online:https://www.itu.int/pub/R-REP-M.2410-2017, accessed 2018-06-29.

[46] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[47] T. M. Nguyen, V. N. Ha, and L. B. Le, "Resource allocation optimization in multi-user multi-cell massive mimo networks considering pilot contamination," *IEEE Access*, vol. 3, pp. 1272–1287, 2015.

[48] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2522–2545, Fourthquarter 2016.

[49] K. Zhu, E. Hossain, and D. Niyato, "Pricing, spectrum sharing, and service selection in two-tier small cell networks: A hierarchical dynamic game approach," *IEEE Transactions on Mobile Computing*, vol. 13, no. 8, pp. 1843–1856, Aug. 2014.

[50] C. Lai, S. B. Chundrigar, S. T. Talat, and H. Chang, "Staying connected on the go : Network sharing based on multioperator traffic inside moving transportation," *IEEE Vehicular Technology Magazine*, vol. 13, no. 1, pp. 97–104, March 2018.

[51] J. Seo and P. Lohan, "Pricing in small cell deployment," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1615–1618, Aug 2016.

[52] L. Wang and G. S. G. S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks; a tutorial," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 271–292, First 2013.

[53] J. Antoniou and A. Pitsillides, "4g converged environment: Modeling network selection as a game," in *2007 16th IST Mobile and Wireless Communications Summit*, July 2007, pp. 1–5.

[54] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[55] I. Malanchini, M. Cesana, and N. Gatti, "Network selection and resource allocation games for wireless access networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp. 2427–2440, Dec 2013.

[56] S. Hamouda, M. Zitoun, and S. Tabbane, "Win win relationship between macrocell and femtocells for spectrum sharing in LTE-A," *IET Communications*, vol. 8, no. 7, pp. 1109–1116, May 2014.

[57] Y. Yi, J. Zhang, Q. Zhang, and T. Jiang, "Spectrum leasing to femto service provider with hybrid access," in *2012 Proceedings IEEE INFOCOM*, Mar. 2012, pp. 1215–1223.

[58] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *2013 Proceedings IEEE INFOCOM*, April 2013, pp. 3303–3308.

[59] T. Fujiwara-Greve, *Non-Cooperative Game Theory*. Tokyo: Springer Japan, 2015, vol. 1.

[60] C. Yang, J. Li, and A. Anpalagan, "Strategic bargaining in wireless networks: basics, opportunities and challenges," *IET Communications*, vol. 8, no. 18, pp. 3435–3450, 2014.

[61] C. Yang, K. Guo, M. Sheng, J. Li, and J. Yue, "Energy-efficient capacity offload to smallcells with interference compensation," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2014, pp. 1745–1749.

[62] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Nash bargaining solution based user association optimization in hetnets," in *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, Jan 2014, pp. 587–592.

[63] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service hetnets using nash bargaining solution," *IEEE Communications Letters*, vol. 18, no. 3, pp. 463–466, March 2014.

[64] D. Niyato, E. Hossain, and Z. Han, "Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: A game-theoretic modeling approach," *IEEE Transactions on Mobile Computing*, vol. 8, no. 8, pp. 1009–1022, Aug 2009.

[65] S. Moon, H. Kim, and Y. Yi, "Brute: Energy-efficient user association in cellular networks from population game perspective," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 663–675, Jan. 2016.

[66] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Communications Letters*, vol. 15, no. 10, pp. 1084–1087, Oct. 2011.

[67] T. Kudo and T. Ohtsuki, "Cell range expansion using distributed q-learning in heterogeneous networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, p. 61, Mar 2013. [Online]. Available: https://doi.org/10.1186/1687-1499-2013-61

[68] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, Feb 2014.

[69] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, February 2014.

[70] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 2457–2461.

[71] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[72] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, May 2017.

[73] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.

[74] N. Saha and R. Vesilo, "An evolutionary game theory approach for joint offloading and interference management in a two-tier hetnet," *IEEE Access*, vol. 6, no. 99, pp. 1807 – 1821, Dec. 2017.

[75] M. Mangili, F. Martignon, S. Paris, and A. Capone, "Bandwidth and cache leasing in wireless information-centric networks: A game-theoretic study," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 679–695, Jan. 2017.

[76] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.

[77] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, "Caching and operator cooperation policies for layered video content delivery," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[78] "A survey on networking games in telecommunications," *Computers and Operations Research*, vol. 33, no. 2, pp. 286 – 311, 2006, game Theory: Numerical Methods and Applications.

[79] Z. Han, D. Niyato, W. Saad, T. Baar, and A. Hjrungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*, 1st ed. New York, NY, USA: Cambridge University Press, 2012.

[80] J. F. Nash, "The bargaining problem," *Econometrica*, vol. 18, no. 2, pp. 155–162, 1950. [Online]. Available: http://www.jstor.org/stable/1907266

[81] K. G. Binmore, *Fun and games : a text on game theory.* Lexington, Mass. : D.C. Heath, 1992, published simultaneously in Canada.

[82] H. J. Peters, *Axiomatic Bargaining Game Theory.* Springer, Netherlands, 1992.

[83] Z. Han, Z. Ji, and K. J. R. Liu, "Fair multiuser channel allocation for ofdma networks using nash bargaining solutions and coalitions," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1366–1376, Aug 2005.

[84] G. Owen, *Game theory*, 2nd ed. New York: Academic Press, 1982.

[85] H. Park and M. van der Schaar, "Bargaining strategies for networked multimedia resource management," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3496–3511, July 2007.

[86] H. Yaiche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 667–678, Oct 2000.

[87] K. D. Lee and V. C. M. Leung, "Fair allocation of subcarrier and power in an ofdma wireless mesh network," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2051–2060, Nov 2006.

[88] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4460080106

[89] R. Mazumdar, L. G. Mason, and C. Douligeris, "Fairness in network optimal flow control: optimality of product forms," *IEEE Transactions on Communications*, vol. 39, no. 5, pp. 775–782, 1991.

[90] C. Yang, J. Li, and A. Anpalagan, "Cooperative bargaining game-theoretic methodology for 5g wireless heterogeneous networks," *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 1, pp. 70–81, 1 2015.

[91] C. G. Yang, J. D. Li, and Z. Tian, "Optimal power control for cognitive radio networks under coupled interference constraints: A cooperative game-theoretic perspective," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1696–1706, 2010.

[92] L. Gao, G. Iosifidis, J. Huang, L. Tassiulas, and D. Li, "Bargaining-based mobile data offloading," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1114–1125, June 2014.

[93] J. M. Smith and G. R. Price, "The logic of animal conflict," *Nature*, vol. 246, no. 1, pp. 15–18, Nov. 1973.

[94] W. H. Sandholm, *Population Games and Evolutionary Dynamics*, 3rd ed. The MIT Press, Nov. 2010.

[95] P. Hammerstein and R. Selten, "Game theory and evolutionary biology," *Handbook of Game Theory with Economic Applications*, vol. 2, pp. 929–993, 1994.

[96] P. D. Taylor and L. B. Jonker, "Evolutionary stable strategies and game dynamics," *Mathematical Biosciences*, vol. 40, no. 1, pp. 145 – 156, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0025556478900779

[97] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006.

[98] H. Tembine, E. Altman, R. El-Azouzi, and Y. Hayel, "Evolutionary games in wireless networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 3, pp. 634–646, June 2010.

[99] M. A. Khan, H. Tembine, and A. V. Vasilakos, "Evolutionary coalitional games: design and challenges in wireless networks," *IEEE Wireless Communications*, vol. 19, no. 2, pp. 50–56, 2012.

[100] L. Xu, H. Fang, and Z. Lin, "Evolutionarily stable opportunistic spectrum access in cognitive radio networks," *IET Communications*, vol. 10, no. 17, pp. 2290–2299, 2016.

[101] C. Xu, M. Sheng, V. S. Varma, T. Q. S. Quek, and J. Li, "Wireless service provider selection and bandwidth resource allocation in multi-tier hcns," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 5108–5124, Dec. 2016.

[102] J. Elias, F. Martignon, L. Chen, and E. Altman, "Joint operator pricing and network selection game in cognitive radio networks: Equilibrium, system dynamics and price of anarchy," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 9, pp. 4576–4589, Nov. 2013.

[103] T. Alpcan, T. Basar, R. Srikant, and E. Altman, "Cdma uplink power control as a noncooperative game," in *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)*, vol. 1, Dec 2001, pp. 197–202 vol.1.

[104] H. A. H. A. Simon, *Models of man social and rational : mathematical essays on rational human behavior in a social setting.* New York: Wiley, 1957.

[105] D. Kahneman, "A perspective on judgment and choice: Mapping bounded rationality," *American Psychologist*, vol. 58, no. 9, pp. 697–720, 2003.

[106] R. Selten, "Bounded rationality," *Journal of Institutional and Theoretical Economics*, vol. 146, no. 4, pp. 649–658, 1990.

[107] ——, "What is bounded rationality?" in *Paper prepared for the Dahlem Conference 1999*, 1999, pp. 1–25.

[108] S. Tan, Y. Wang, Y. Chen, and Z. Wang, "Evolutionary dynamics of collective behavior selection and drift: Flocking, collapse, and oscillation," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1694–1705, July 2017.

[109] P. Semasinghe, E. Hossain, and K. Zhu, "An evolutionary game for distributed resource allocation in self-organizing small cells," *IEEE Transactions on Mobile Computing*, vol. 14, no. 2, pp. 274–287, Feb. 2015.

[110] S. S. M. W. Hirsch and R. L. Devaney, *Differential Equations, Dynamical Systems and An Introduction To Chaos.* Elsevier Academic Press, 2003.

[111] "Relaxation heuristics for a generalized assignment problem," *European Journal of Operational Research*, vol. 91, no. 3, pp. 600 – 610, 1996.

[112] D. Bertsekas, *Nonlinear Programming*, ser. Athena scientific optimization and computation series. Athena Scientific, 2016. [Online]. Available: https://books.google.com.au/books?id=TwOujgEACAAJ

[113] "Exact algorithms for the joint object placement and request routing problem in content distribution networks," *Computers and Operations Research*, vol. 35, no. 12, pp. 3860 – 3884, 2008, part Special Issue: Telecommunications Network Engineering.

[114] R. Nasri, S. Affes, and A. Stephene, "Combined interference cancellation and avoidance over the downlink of spectrum-sharing lte hetnet," in *2015 IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB)*, 2015, pp. 1–6.

[115] S. Hamouda, I. B. Chaabane, and S. Tabbane, "Cooperative bandwidth sharing for relaying in lte-advanced using game theory," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2306–2317, 2015.

[116] B. Yang, G. Mao, M. Ding, X. Ge, and X. Tao, "Dense small cell networks: From noise-limited to dense interference-limited," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4262–4277, May 2018.

[117] W. Liu, S. Han, C. Yang, and C. Sun, "Massive mimo or small cell network: Who is more energy efficient?" in *2013 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, April 2013, pp. 24–29.

[118] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 678–700, 2013.

[119] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-qos-aware fair scheduling for lte," in *2011 IEEE 73rd Vehicular Technology Conference*, 2011.

[120] K. Sandrasegaran, H. A. M. Ramli, and R. Basukala, "Delay-prioritized scheduling for real time traffic in 3gpp lte system," in *2010 IEEE Wireless Communication and Networking Conference*, 2010, pp. 1–6.

[121] H. A. M. Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, M. Xue, and C.-C. Lin, "Resource allocation technique for video streaming applications in the lte system," in *The 19th Annual Wireless and Optical Communications Conference (WOCC 2010)*, 2010, pp. 1–5.

[122] L. Wang, G. Tyson, J. Kangasharju, and J. Crowcroft, "Milking the cache cow with fairness in mind," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2686–2700, Oct 2017.

[123] D. Mackey, "Numerical methods for finding the roots of a function," 2015, online:http://www.maths.dit.ie/~dmackey/lectures/Roots.pdf, accessed 2018-06-29.

[124] M. Held and R. M. Karp, "The traveling-salesman problem and minimum spanning trees," *Operations Research*, vol. 18, no. 6, pp. 1138–1162, 1970.

[125] ——, "The traveling-salesman problem and minimum spanning trees: Part ii," *Math. Program.*, vol. 1, no. 1, pp. 6–25, Dec. 1971. [Online]. Available: https://doi.org/10.1007/BF01584070

[126] M. L. Fisher, "The lagrangian relaxation method for solving integer programming problems," *Management Science*, vol. 50, no. 12_supplement, pp. 1861–1871, 2004.

[127] V. Y. F. Tan, "Simplified proof of slater's theorem for strong duality," 2015, online:https://www.ece.nus.edu.sg/stfpage/vtan/ee5138/slater.pdf, accessed 2018-06-29.

[128] R. Deng, Y. Zhang, S. He, J. Chen, and X. Shen, "Maximizing network utility of rechargeable sensor networks with spatiotemporally coupled constraints," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1307–1319, May 2016.

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

[130] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University*, 2015.

[131] S. Guo, X. Zhou, S. Xiao, and M. Sun, "Fairness-aware energy-efficient resource allocation in d2d communication networks," *IEEE Systems Journal*, pp. 1–12, 2018.

[132] S. Xu, R. Li, and Q. Yang, "Improved genetic algorithm based intelligent resource allocation in 5g ultra dense networks," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[133] A. J. Roumeliotis, S. Vassaki, and A. D. Panagopoulos, "Qos-driven power and time allocation scheme for spectrum leasing in overlay cognitive radio networks," *IET Communications*, vol. 12, no. 6, pp. 688–695, 2018.

[134] R. Harwahyu, R. G. Cheng, C. H. Wei, and R. F. Sari, "Optimization of random access channel in nb-iot," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 391–402, Feb 2018.

[135] B. Wu, L. Tong, Y. Chen, and L. He, "New methods in multibaseline polarimetric sar interferometry coherence optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2016–2020, Oct 2015.

[136] Y. R. Tsai and H. A. Wei, "Quality-balanced user clustering schemes for non-orthogonal multiple access systems," *IEEE Communications Letters*, vol. 22, no. 1, pp. 113–116, Jan 2018.

[137] D. Maringer, *Portfolio management with heuristic optimization / by Dietmar Maringer.*, ser. Advances in computational management science ; v. 8. Dordrecht: Springer, 2005.

[138] "Heuristic search," in *Heuristic Search Theory and Applications*, S. Edelkamp and S. Schrödl, Eds. San Francisco: Morgan Kaufmann, 2012, p. iv. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123725127000353

[139] C. F. Gerald, *Applied numerical analysis / Curtis F. Gerald, Patrick O. Wheatley.*, 4th ed., ser. World student series. Reading, Mass: Addison-Wesley Pub. Co., 1989.

[140] A. Ralston, *A first course in numerical analysis.*, ser. International series in pure and applied mathematics. New York: McGraw-Hill, 1965.

[141] G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part i — convex underestimating problems," *Mathematical Programming*, vol. 10, no. 1, pp. 147–175, Dec 1976. [Online]. Available: https://doi.org/10.1007/BF01580665

[142] L. Liberti and C. C. Pantelides, "An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms," *Journal of Global Optimization*, vol. 36, no. 2, pp. 161–189, Oct 2006.

[143] C. Y. Chang, S. Martínez, and J. Cortés, "Convex relaxation for mixed-integer optimal power flow problems," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2017, pp. 307–314.

[144] Z. Zhu, J. Peng, X. Gu, H. Li, K. Liu, Z. Zhou, and W. Liu, "Fair resource allocation for system throughput maximization in mobile edge computing," *IEEE Access*, vol. 6, pp. 5332–5340, 2018.

[145] M. Petrik and S. Zilberstein, "A bilinear programming approach for multiagent planning," 2014.

[146] O. Mangasarian, "Mathematical programming in data mining," *Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 183–201, Jun 1997. [Online]. Available: https://doi.org/10.1023/A:1009735908398

[147] L. Aitchison, N. Corradi, and P. Latham, "Zipf's law arises naturally when there are underlying, unobserved variables," *PLoS Computational Biology*, vol. 12, no. 12, 2016. [Online]. Available: http://search.proquest.com/docview/1858867704/

[148] M. Hajimirsadeghi, N. B. Mandayam, and A. Reznik, "Joint caching and pricing strategies for popular content in information centric networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 654–667, March 2017.