Clustering Nature of Base Stations and Traffic Demands in Cellular Networks and the Corresponding Caching and Multicast Strategies

Yifan Zhou

ANNÉE 2018

THÈSE EN COTUTELLE

pour obtenir le grade de Docteur délivré par

UNIVERSITE DU ZHEJIANG
Spécialité : Ingénierie de l'information et des Communications
Ecole doctorale : 0810

****

CENTRALESUPÉLEC
sous le sceau de l'Université Bretagne Loire
Spécialité : Télécommunications
Ecole doctorale 601 MathSTIC

présentée par

Yifan ZHOU

Préparée à
Collège des sciences de l'information et Ingénierie électronique
UMR 6164 - IETR - Institut d'Electronique et de Télécommunications de Rennes

---

Titre de la thèse

Clustering Nature of Base Stations and Traffic Demands in Cellular Networks and the Corresponding Caching and Multicast Strategies

Thèse soutenue à Rennes
le 3 juillet 2018

devant le jury composé de :

Youping ZHAO
Professeur, Université de Beijing Jiao Tong, Chine / *rapporteur*

Aline Carneiro Viana
Chargée de recherche, HDR, INRIA-Saclay / *rapporteur*

Jacques PALICOT
Professeur, CentraleSupélec / *examinateur*

Xianfu CHEN
Docteur-Ingénieur, VTT Finlande / *examinateur*

Yves LOUET
Professeur, CentraleSupélec / *directeur de thèse*

Honggang ZHANG
Professeur, Université de Zhejiang, Chine / *co-directeur de thèse*

Institutional Acknowledgements

This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference No. ANR-10-LABX-07-01. The authors would also like to thank the Region Bretagne, France, for its support of this work.

Acknowledgements

First of all, I would like to express my sincere gratitude to my thesis advisors Yves Louët, Honggang Zhang and Jacques Palicot for giving me the chance to pursuit the joint doctoral degree between CentraleSupélec and Zhejiang university. I'm greatly honoured to have lived and studied in Campus Rennes for one year. During my stay, the colleagues and friends impressed me so much with their warm helps and patient guidances, not to mention the productive environment and French humors. I would also like to thank Christophe Moy, Carlos Bader, Karine Bernard, Amor Nakfha and Pascal Cotret for their helps and regards.

Moreover, I had the pleasure to work with a very friendly team. I would like to thank Navikkumar Modi, Quentin Bodinier, Remi Bonnefoi, Mouna Ben Mabrouk, Haïfa Farès, Marwa Chaffi, Vincent Gouldieff, Lilian Besson, Muhammad Abdul Wahab and Rami Othman for their kindness and company. During my stay in Rennes, I have also made acquaintance with many Chinese friends. I would to thank Qipeng Song, Jialong Duan, Yue Li and Hao Jiang for their understanding and suggestions.

Finally, I would like to thank my family for their enduring love and support.

Abstract

Traditional cellular networks have evolved from the first generation of analog communications to the current fourth generation of digital communications where iteratively enhanced physical layer technologies have greatly increased the network capacity. According to Shannon's theory, the technical gains brought by physical layer has gradually become saturated, which cannot match the rapid increase of user traffic demand in current mobile internet era, thus calls for another path of evolution, i.e., digging into the traffic demand of mobile users. In recent years, the academic communities have begun to use the real data to analyze the infrastructure deployment of wireless networks and the traffic demand of mobile users, in order to make benefits from the underlying statistical patterns. At the same time, along with the recent rise of machine learning technics, data-driven service is considered as the next economic growth point. Thus the industry is putting more and more attention on data accumulation and knowledge mining related services and telecommunication operators are coming to realize the increasing importance of the recorded data from their own networks. Therefore, the real-data-driven technology advancement is considered as a promising direction for the next evolution of cellular networks.

In this thesis, we firstly gave a comprehensive review of the state-of-the-art real data measurements in Chapter 2 which not only sheds light on the importance of real data analysis, but also paves way for its reasonable usage to improve the service performance of cellular networks. From the survey, we concluded that there exhibits a periodic pattern for the temporal traffic assumption of large coverage area in cellular networks, while for single cell, a heavy-tailed distribution is widespread across the temporal and spatial characterization. Furthermore, this imbalance phenomenon emerges more significantly in the call duration, request arrivals and content preference of mobile users.

Then, based on a large amount of real data collected from on-operating cellular networks, we conducted a large-scale identification on spatial modeling of base stations (BSs) in Chapter 3. According to the fitting results, we verified the inaccuracy of Poisson distribution for BS locations, and uncovered the clustering nature of BS deployment in cellular networks. However,

although typical clustering models have improved the modeling accuracy but are still not qualified to accurately reproduce the practical BSs deployment, which leads to the spatial density characterization of BS.

In Chapter 4, we tried to characterize the density of BS deployment and traffic demand, in both spatial domain and temporal dimensions. In accordance with the heavy-tailed phenomenons in Chapter 2, we found that the α-Stable distribution is the most accurate model for the spatial densities of BSs and traffic consumption, between which a linear dependence is revealed through real data examination. Moreover, the accuracies of power-law and lognormal distributions for the packet length and inter-arrival time of user requests are verified, respectively, which convincingly leads to the α-Stable distribution of temporally aggregated traffic volume on BS level.

To make benefit from the findings in previous chapters, we proposed a cooperative caching strategy based on the spatial clustering of BSs and a dynamic unicast/multicast strategy based on the temporal aggregation of content requests in Chapter 5. According to the theoretical and simulation results, we found that the proposed 'Caching as a Cluster' strategy can significantly reduce the average delay of users especially in the inhomogeneous BS deployment scenario, and the dynamic unicast/multicast strategy can not only reduce the average latency of content requests but also diminish the average power consumption of BSs especially under the bursty request arrival patterns.

To implement the massive real data analyses and dynamic serving mechanisms aforementioned, we proposed an intelligent SDN-based centralized architecture within cellular networks in Chapter 5. With the introduction of an intelligence center, the brand new architecture is able to trace the demand variations in real time, thus simultaneously satisfy the operational requirements of the entire network and QoEs of all users by deploying flexible and efficient algorithms upon it.

Conclusively, in this thesis, we uncovered the clustering nature of cellular networks in different dimensions, and proposed corresponding service strategies to tackle the clustering challenge and utilize them for efficiency improvement.

Contents

Abbreviations

3GPP	3rd Generation Partnership Project
BS	Base Station
BSC	Base Station Controllers
BSS	Base Station Subsystem
BTS	Base Transceiving Stations
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CDN	Content Delivery Networks
CDR	Call Detail Records
CN	Core Networks
EPC	Enhanced Packet Core
GGSN	Gateway GPRS support node
GP	generalized Pareto
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HLR	Home Location Register
HSPA	High-Speed Packet Access
HTTP	HyperText Transfer Protocol
LTE	Long-Term Evolution
MCP	Matern Cluster Process
MIM	Mobile Instant Message
MLE	Maximum Likelihood Estimation
MSC	Mobile Switching Centers
NSS	Network and Switching Subsystem
OTT	Over the Top
PDF	Probability Density Function
PGW	Packet Gateways
PHCP	Poisson Hardcore Process
PPP	Poisson Point Process
QoE	Quality of Experience

RAN	Radio Access Network
RMSE	Root Mean Square Error
RNC	Radio Network Controllers
RRC	Radio Resource Control
SDN	Software Defined Networking
SGSN	Serving GPRS Support Node
SGW	Serving Gateways
SINR	Signal to Interference Noise Ratio
SIR	Signal to Interference Ratio
SMS	Short Messages Service
TCP	Thomas Cluster Process
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	Universal Terrestrial Radio Access Network
VLR	Visitor Location Register

Résumé en Français

Les réseaux cellulaires traditionnels ont évolué de la première génération à base de communications analogiques à la quatrième génération de communications numériques, où les technologies de la couche physique se sont améliorées de façon à considérablement augmenter la capacité du réseau. Selon la théorie de Shannon, les gains apportés par la couche physique vont être progressivement saturés, ce qui ne permet pas de suivre l'augmentation rapide de la demande de trafic des utilisateurs dans l'ère actuelle de l'Internet mobile. Ces dernières années, les communautés universitaires ont commencé à utiliser les données réelles pour analyser le déploiement de l'infrastructure des réseaux sans fil et la demande de trafic des utilisateurs mobiles, afin de tirer parti des modèles statistiques sous-jacents. Dans le même temps, avec la récente montée en puissance des techniques d'apprentissage automatique, l'analyse de données est considérée comme le prochain verrou de croissance économique. Ainsi, l'industrie accorde de plus en plus d'attention à l'accumulation de données et les services liés à l'exploitation des connaissances et les opérateurs de télécommunications commencent à prendre conscience de l'importance croissante des données enregistrées à partir de leurs propres réseaux. Par conséquent, le progrès technologique axé sur les données réelles est considéré comme une orientation prometteuse pour la prochaine évolution des réseaux cellulaires.

Dans cette thèse, nous avons tout d'abord passé en revue les mesures de données réelles issues d'opérateurs dans le chapitre 2 qui non seulement mettent en lumière l'importance de l'analyse de données réelles, mais ouvrent également la voie à la possibilité d'améliorer les performances de service des réseaux cellulaires. Dans cette analyse, nous avons conclu qu'il existe un modèle périodique pour l'hypothèse de trafic temporel de grande zone de couverture dans les réseaux cellulaires, tandis que pour une cellule unique, une distribution de type « heavy tail » caractérise la caractérisation temporelle et spatiale. Ce phénomène de déséquilibre apparaît plus significativement dans la durée de l'appel, d ans les demandes d'arrivée et la préférence de contenu des utilisateurs mobiles.

Ensuite, sur la base d'une grande quantité de données réelles collectées à partir des réseaux cellulaires en fonctionnement, nous avons effectué une identification à grande échelle sur la modélisation spatiale des stations de base (BS) dans le chapitre 3. Selon les résultats de cet

ajustement, nous avons vérifié l'inexactitude de la distribution de Poisson pour les emplacements de BS, et mise en évidence le regroupement dans le déploiement de BS dans les réseaux cellulaires. Cependant, bien que les modèles de regroupement typiques aient amélioré la précision de la modélisation, ils ne sont pas encore valides pour reproduire fidèlement le déploiement pratique des BS, ce qui conduit à la caractérisation de la densité spatiale de BS.

Dans le chapitre 4, nous avons caractérisé la densité du déploiement BS et de la demande de trafic, à la fois dans le domaine spatial et mais aussi dans le domaine temporel. En accord avec les phénomènes de type « heavy tail » du chapitre 2, nous avons trouvé que la distribution -Stable était le modèle le plus précis pour les densités spatiales des BS et de la distribution de trafic, entre lesquelles une dépendance linéaire est révélée en analysant les données réelles. De plus, les exactitudes des lois de puissance et lognormales pour la longueur des paquets et l'heure d'arrivée des requêtes des utilisateurs sont vérifiées respectivement, ce qui conduit de manière convaincante à la distribution -Stable du volume de trafic agrégé temporellement au niveau BS.

Pour tirer profit des résultats des chapitres précédents, nous avons proposé une stratégie de type « caching » coopérative basée sur le clustering spatial des BS et une stratégie unicast/multicast dynamique basée sur l'agrégation temporelle des requêtes de contenus, développé dans le chapitre 5. Selon les résultats théoriques et de simulation, nous avons mis en évidence que la stratégie de mise en cache proposée en clusters peut réduire considérablement le retard moyen des utilisateurs, en particulier dans le scénario de déploiement BS non homogène, et la stratégie dynamique unicast/multicast peut non seulement réduire la latence moyenne des demandes de contenu, mais aussi diminuer la consommation d'énergie moyenne des stations de base, en particulier sous les modèles d'arrivée de demande de contenus.

Pour mettre en œuvre les analyses massives de données réelles et les mécanismes de desserte dynamiques susmentionnés, nous avons proposé une architecture centralisée de type SDN (Software Defined Network) intelligent au sein des réseaux cellulaires dans le chapitre 5. Avec l'introduction d'un moteur de décision, la nouvelle architecture est capable de tracer les variations de la demande en temps réel, répondant ainsi simultanément aux exigences opérationnelles de l'ensemble du réseau et des QoE de tous les utilisateurs en déployant des algorithmes flexibles et efficaces.

En conclusion, dans cette thèse, nous avons mise en évidence la nature de regroupement des stations de base des réseaux cellulaires dans différentes dimensions, et proposé des stratégies de service correspondant pour s'attaquer au défi de regroupement et les utiliser pour l'amélioration de l'efficacité.

Chapter 1

Introduction

Contents

1.1 Background

With the world-spread popularity of smart devices, including phones, tablets and numerous virtual reality devices, the mobile user traffic demand is increasing exponentially, unlike the traditional voice service generations ago. According to the latest Cisco visual networking index white paper, there will be another 8-fold increase of mobile traffic (i.e., from 3.7 to 30.6 exabytes per month from 2015 to 2020) [23]. Driven by this tremendous data consumption and higher quality-of-experience (QoE) requirement, research groups including academics and industrials, devote themselves to upgrade the cellular networks by looking into the user demand dynamics and proposing brand new technologies.

On one hand, with the amazing development of big data technologies these years, operators are coming to realize the increasing importance of recorded data from their own networks. Thus, more and more real data are put into investigation by the operators themselves or contributed to related researchers aiming to reveal the inherit patterns of mobile user activities and mobile traffic. On the other hand, along with the 3rd generation partnership project (3GPP) releases, many researchers are working on new architecture or algorithm to increase the network capacity and improving the QoE of mobile subscribers. However, the effectiveness or efficiency of these new kinds of technologies in cellular networks need to be verified based on extensive real data measurements.

To make full use of these methodologies in two different directions, the best way is to combine these two methods together smoothly. That's to say, based on the real data examination part, large-scale identification can give rise to accurate and reasonable characterizing models which can be adopted in the technology verification part. Actually, the research topics of real data in cellular networks include traffic, base stations (BSs) and mobile users, along with their spatial and temporal distributions. Therefore, these realistic mathematical models can be adopted as preliminary assumptions for academical and industrial research, such as the recently popular stochastic geometry domain which is based on the homogeneous poisson point process (PPP) distribution of BSs or mobile users.

Furthermore, besides the verification benefits of realistic network models, what makes real data so important is that the dynamics of users' traffic patterns or the so-called evolution of cellular networks can only be reflected by real network measurements. For example, the exponential distribution is usually adopted as the arrival pattern of user traffic in 3GPP's protocol design. However, due to the explosively growing user number and traffic amount these years, the bursty nature of cellular networks is becoming more and more universal which makes the exponential assumption not reliable any more.

1.2 Research Topics

This thesis focuses on the unveiling of clustering nature in cellular networks, and evaluate their potential impacts on the service performance.

Firstly in Chapter 2, we gave a comprehensive review on the statistical characteristics of traffic demand in cellular networks on different dimensions (space, time and content). In detail, we reviewed the state-of-the-art temporal analysis of cellular traffic consumption, on both macro view (aggregated traffic on cells or upper level) and micro view (individual traffic of user or application level). According to related works, the traffic demand in cellular networks exhibits universal clustering nature on different time scales and on different aggregation levels. In parallel, we also introduced the spatial examination of traffic demand, and it also expresses significant level of aggregation effect. At last, besides the temporal and spatial dimensions, we also revealed the heavy-tailed property of traffic demand on content dimension. That's to say, mobile users tend to request the same content and the popularity distribution of possible requested contents is distinctly unbalanced. Furthermore, it's important to know that the statistical features of cellular traffic basically result from the usage pattern of mobile users, on all three different dimensions. Therefore, we can conclude that the mobile users in cellular networks are also clusteringly distributed, their temporal usages of traffic are also aggregated and their content preferences are more or less similarly concentrated.

After that, based on a large amount of real data from on-operating cellular networks, we investigated the spatial distribution of BSs. In detail, the BS locations are examined firstly to check the traditional assumptions, such as the hexagonal placement and homogeneous PPP. According to the fitting results, the widely adopted spatial models are inappropriate for the real deployment. Therefore, we conducted a large-scale comprehensive identification of BS spatial distribution in cellular networks, considering a variety of popular two-dimensional point processes including repulsive and attractive ones, and tried to figure out the most accurate spatial model for both macro and micro BSs. Unfortunately, based on those extensive real data, all of these well understood point processes are not qualified enough to characterize the realistic BSs distribution according to the random verifications and multiple performance metrics. Details can be found in Chapter 3.

Since the two-dimensional point processes are not qualified to characterize the realistic BSs' spatial distribution, we turned to the one-dimensional spatial density distribution of BSs in cellular networks. Based on the same set of real data, we conducted a general fitting process to find the most accurate statistical distribution of BSs' spatial density. After random area sampling in the under-investigated cellular networks, we choose several popular probability density functions (PDFs), including power-law, Weibull, log-normal and α-Stable distribution along with the traditional Poisson distribution as candidates for the fitting procedure. According to the root mean square error (RMSE) performance metric, we found that the α-Stable distribution can best reflects the clustering nature of BS deployment in cellular networks. Besides, along with the spatial distribution of BS, we tried to connect the dots between the data traffic and the mobile user distribution. After examining the spatial distribution of these three dynamics, we found out that the spatial density of users, traffic and BSs are linearly growing with each other, revealing a trinity-like entangled phenomenon. Details can be found in Chapter 4.

After revealing the clustering nature of cellular networks in different dimensions, we wanted to make good use of these unconventional properties. For example, by combining the content preference of mobile users and the spatial clustering of BSs, we investigated the distributed probabilistic caching strategy in the clustered heterogeneous cellular networks, based on the cooperation between BSs in the same spatial group. Assuming that nearby BS pairs share a limited interchangeable bandwidth, the 'Caching as a Cluster' type of BS collaboration helps to reduce the overall content delivery latency of mobile users in BS caching scenarios, as been illustrated in the first part of Chapter 5.

Besides, considering the temporal burstiness of traffic demand and spatial clustering of BSs, there is a chance for the introduction of BS sleeping strategies into cellular networks to improve the overall energy efficiency. Actually, the basic principle is that the default traffic demand of BS which encounters a low-load situation can be served by the nearby ones with tolerable QoE and inexpensive cost because the BSs are spatial clusteringly distributed. Intuitively, in this

cluster scenario, the overall performance will be better than that of the homogeneous case. In fact, this topic has been widely studied in the green communication fields in these year, and related works can be found in Chapter 2. As we focus on the caching and multicast, so the BS sleeping is not covered in this thesis and will be considered as future works.

On the other hand, by combining the content preference of mobile users and the temporal bursty nature of traffic demand, we can adopt the multicast technic in cellular networks considering its broadcasting nature, in order to diminish the number of wireless transmissions and ensure the QoE requirement of users at the mean time. In this case, the spatially clustering nature of mobile user is another necessity to make the multicast procedure practical and beneficial. Thus, three dimensional aggregation properties are all considered in this scenario, which can make a noticeable difference on the capacity improvement comparing to the traditional broadcasting technics in wireless networks. All these technics based on different combination of clustering dimension are introduced in the second part of Chapter 5.

Above all, given these clustering nature of mobile users, traffic demand and BSs on different dimensions (i.e. temporal, spatial and content), which are certificated by massive real data, a new paradigm of service procedure can be explored. That's to say, firstly analyzing the statistical characteristics of both mobile user and fixed infrastructure by collecting dynamic registration or static deployment records, then monitoring the overall traffic demand by collecting the online content requests distributively, finally go through a centralized processor with holistic information to provide a globally optimized service solution. In fact, this kind of service paradigm needs high-performance data processing technics like data mining, traffic prediction methods, along with high-capacity central controller for fast and reliable instruction delivery and this task provides a appropriate scenario for big data technology and software defined networking (SDN) paradigm. This new centralized architecture along with two use cases (i.e., caching and multicast) are introduced in the third part of Chapter 5.

1.3 Contributions

The contributions of this thesis can be concluded to several parts as follows.

Firstly, we gave a comprehensive review of the state-of-the-art real data measurement in cellular networks. Specifically, the related works cover those data analysis in mobile users, traffic aggregation and network infrastructure, on different dimensions including temporal, spatial and contents, also in different perspectives like resource allocation, QoE and energy efficiency. Most of these measurement results reveal kind of clustering nature in cellular networks.

Secondly, we conducted a large-scale identification on the spatial distribution of BSs in cellular networks, separately on typical urban, rural sample regions and on numbers of randomly

selected regions. By choosing most popular point processes as fitting candidates, we tried to find out the most appropriate spatial model for BS locations. However, all these well understood two-dimensional point process models are rejected by the large amount of real data according to either classical statistical or coverage probability performance metrics. After all, instead of pursuing a two-dimensional point process model for spatial distribution, we turned to the one-dimensional BS density distribution characterization and find out that α-Stable is the most accurate statistical distribution for BS and traffic demand density reflecting the explicit clustering nature of infrastructure deployment in cellular networks.

Thirdly, to make full use of the clustering nature of user requested contents and the spatial distribution of BSs, we introduced a collaborative caching strategy in clustered cellular networks where small BSs within the same cluster can exchange contents with each other constrained by a given bandwidth. According to the simulation results, we found that the cooperation scheme can reduce the average content delivery latency of mobile users significantly. Besides, we investigated the impact of BS density on the caching performance, and provided a overall latency evaluation of the clusteringly distributed cellular networks.

Fourthly, considering the clustering nature on time and content dimension of mobile users traffic demand, we proposed to adopt the multicast technics into cellular networks. Before diving into the utilization of temporal burst nature of content requests, we firstly derived the theoretical result on the average latency of user requests in the Poisson arrival scenario, and based on which the unicast/multicast hybrid transmission strategy is proposed. According to results from both theoretical derivation and simulation verification, we found the introduction of multicast can not only solve the problem that the average latency of user request increases indefinitely with the arrival rate under congestion, but also is able to reduce the average power consumption of BS in the hybrid transmission process due to the natural advantage of one-to-all scheme.

Fifthly, we put forward an intelligent-SDN (Software Defined Networking) based centralized architecture within cellular networks. Through the collection of large amount of real-time recording, and the intelligent and efficient algorithms, the proposed architecture provides computing and storage resources for the cooperative cache strategy of the access network and the unicast/multicast hybrid transmission strategy at the BS described in previous chapters, in order to achieve highly efficient and real-time service attributes.

Conclusively, this thesis investigated the clustering nature of cellular networks in different dimensions and utilized them to improve the service efficiency of several well known communication technologies. In addition to the theoretical and simulating analysis, we put forward a promising intelligent controlling architecture packed with machine learning algorithms and SDN technics to make full use of the clustering nature in cellular networks.

The presentation of this thesis is organized as follows, Chapter 2 firstly gives a comprehensive introduction to the related works in cellular networks traffic measurements; Chapter 3 presents a specific investigation on the spatial distribution of BSs based on two-dimensional point process modeling. Furthermore, the α-Stable distribution is introduced to model the spatial density of BS, traffic demand, and temporal aggregation effect in Chapter 4. After that, the probabilistic caching strategy based on nearby BSs collaboration is analyzed in Chapter 5, along with which the multicast technics utilizing temporal clustering of content requests are also proposed. Above all, after revealing the statistical results and analyzing the technical potentials, an overall intelligent controlling architecture is proposed in the last part of Chapter 5. After all, the conclusion and future works are given in Chapter 6.

Chapter 2

Real Data Measurements in Cellular Networks: Unveiling the Statue

Contents

This chapter will give a comprehensive review on the real data measurements in cellular networks, including analysis of mobile users, traffic demand, and BSs. According to the corresponding characteristics of each considered subject, the introduction is presented on different dimensions, such as temporal dynamics, spatial distributions or content classification.

2.1 Introduction

After the description of the network architecture, this section will display a broad picture of the real data measurements in cellular networks, focusing on explaining why this field rises as

a popular research topic, how important is the real data for the cellular networks research and what would be its benefits for the performance analysis.

2.1.1 Cellular Networks Architecture

Since their invention, cellular networks have been designed as a modular architecture that allows inter-operability among different generations across diverse technologies and distinct requirements. During the past decades, from a functional point of view, the overall structure of cellular networks has been unchanged and physical entities remain grouped into two domains: radio access network (RAN) and core network (CN) domains. The RAN domain provides users with radio resources to access the CN domain (uplink or downlink), while the latter is responsible for the management of services, including the establishment, termination and QoS based parameter reconfiguration. Fig. 2.1 outlines these domains across 2G, 3G, and 4G networks according to the corresponding major standards, i.e., the global system for mobile communications (GSM), the universal mobile telecommunications system (UMTS) and the long-term evolution (LTE), respectively [64].

2G GSM Networks

The RAN domain in 2G networks is named as BS subsystem (BSS). It consists of base transceiving stations (BTS) and BS controllers (BSC). As BTS is responsible for radio transmissions and receptions along with some physical layer processings, a BSC is in charge of a group of BTSs. Moreover, BSC is responsible for the management of radio resources, paging and handover procedures under its coverage area. On the other hand, the CN domain referred to as network and switching subsystem (NSS), only performs circuit-switched function, and it is usually formed by mobile switching centers (MSC) and gateway mobile switching centers which is responsible for voice call control, user equipment (UE) registrations and mobility management. Besides, several major databases useful for managing customers are also included in the CN domain, i.e., the home location register (HLR) which stores the detailed description of registered subscribers, the visitor location register (VLR) which helps for the roaming precedure, the authentication center dealing with the authentication and encryption procedures, and the equipment identity register storing the unique identification of each subscriber.

3G UMTS Networks

The universal terrestrial radio access network (UTRAN) represents the RAN part in UMTS networks and it is composed of NodeBs and radio network controllers (RNC) that match up with BTSs and BSCs in GSM networks. On the other hand, the CN, is divided into two parts:

the circuit-switched and packet-switched domains, representing practically a combination of the GSM NSS, and the general packet radio service (GPRS) backbone. We remark that GPRS is a technology between 2G and 3G cellular networks, that provides mobile data services with data rates of a few kbps. The PS domain consists of the serving GPRS Support Node (SGSN) and the gateway GPRS support node (GGSN), responsible for handling packet connections of UEs, security functionalities and mobility management functions as well as data routing.

4G LTE Networks

In contrast to UMTS systems, LTE networks are designed to provide only PS services. The RAN, or so-called Enhanced-UTRAN in 4G era, is only formed by interconnected BSs called eNodeBs, without centralized controlling entities, which is opposed to its preceding technologies. Similarly, the eNodeB takes responsibility for radio-related functions and it is directly connected to the core network, which is referred to as enhanced packet core (EPC). The EPC, that is responsible for the overall control of UEs, includes serving gateways (SGW), packet gateways (PGW) who manages data packets routing and forwarding, as well as network address allocations, and mobility management entities (MME) performing connection management. Additionally, by cooperating with the following other entities: home subscriber server, enhanced serving mobile location center, and gateway mobile location center, the MME completes mobility-related and authentication-related tasks. Finally, the EPC also includes a policy control and charging rules function entity orchestrating policy and how control decision makings.

For the billing and inter-operator accounting procedures, a set of logical charging functions are implemented in the network. Specifically, these elements collect network resource usages of each customer and implement following functions: the charging trigger function (CTF), which generates charging events based on the observation of network resource usages; the charging data function, which receives charging events from the CTF to construct call detail records (CDR), providing for each user reports concerning his communications; and the charging gateway function, responsible for validating, reformatting and storing CDRs before sending them to the billing domain [71].

After introducing the overall architecture of cellular networks briefly, following subsections will discuss the real data measurements and their underlying applications in literature.

2.1.2 The Importance of Real Data in Cellular Networks Research

As an important record to reflect the operation of cellular networks, measured real data play an indispensable role in the analysis of network deployment and service efficiency. In recent years, the academic communities have begun to use the real data to analyze the deployment

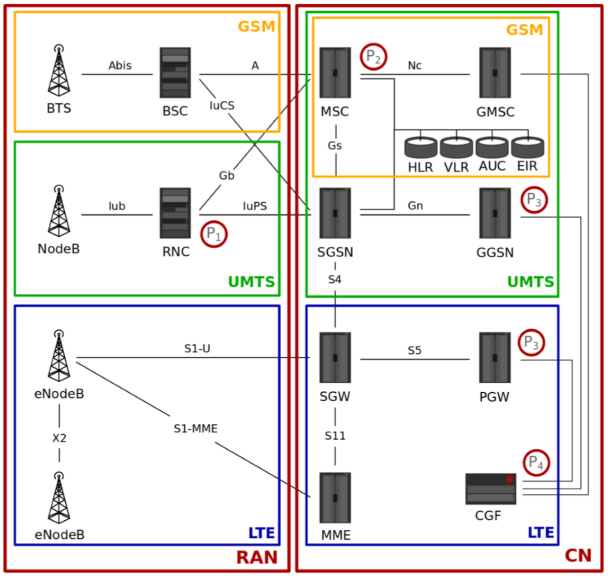


FIGURE 2.1: The architecture of typical cellular networks with different generations of technologies.

characteristics of cellular networks and service characteristics. At the same time, along with the recent rise of machine learning boom, data-driven service is considered as the next economic growth point, thus the industry is putting more and more attention on data accumulation and knowledge mining related services. More importantly, as the source of real data, operators expect this advantage serves for the business transformation, that is, from a communication provider to a content provider, in order to satisfy the more extensive online demands of mobile users directly. Due to various necessities and the rise of many big data technologies, real data analysis has become a hot topic in cellular networks research. Above all, what kind of potential can real data provide for the performance improvement of cellular networks? In which way can it help researchers to better penetrate into the essential attributes of the system? Here we examine the importance of real data in cellular networks research from three aspects.

Real Data as a Source

As a result of the rapid development during the last decade, cellular networks are becoming more and more complicated, as introduced in the last subsection. Therefore, in order to provide an insight view of the internal structure, real data measurement emerges as an accurate and inexhaustible way as database and processing technologies keep upgrading.

Besides, owing to the overwhelming popularity of smart phones in these years, cellular networks are able to provide an all-weather recording of nearly every subscriber around the world including the locations and communication activities [13] which are usually used for billing purposes only. From this point of view, the real data from global cellular networks may not only provide insights into the system itself, but can also make benefits for many other fields, such as human

dynamics [85] [90], city planning [72], traffic prediction [13] and so on. Regarding the related progressive researches, cellular networks actually have been the main contributor of real data worldwide and are promising to reach a general data-fetching platform for inter-field scientific collaboration.

Return to cellular networks analysis itself, actually, the telecommunication operators have been volunteers long time ago to use real data from their own networks to improve the service quality. For example, based on the signal strength across the coverage area and the outage probabilities of different cells, the operators decide where to deploy new BSs and how much is the appropriate transmit power. However, these kinds of usage that lacks overall planning and fundamental principle are not good enough to make full use of real data from cellular networks. As an example, authors in [76] made use of worldwide real data consists of fine-grained RRC (Radio Resource Control) dynamics and provided a detailed analysis of the RRC inter-state timeout setting which significantly contributes to the average latency of user data transmission.

Conclusively, real data in cellular networks can not only be a reliable and active source for problem detection technically, but also provide an accurate and comprehensive source for pattern discovery theoretically.

Real Data as a Standard

Driven by the ever-growing traffic volume and high QoS requirement in cellular networks, a lot of new architectures and technologies have been proposed in recent years claiming that each of them provides more or less performance improvement. However, many propositions of algorithms or methods are merely based on numerical simulations but do not involve any real scenario examination.

Firstly, the performance evaluation solely based on the simulation results is not a wise choice for real network deployment, because it probably underestimates the complexity of actual communication environment. Furthermore, the scale of applicability for a specific newly proposed technique is difficult to predefined, i.e., maybe it's applicable in a single cell but causes unforeseen severe performance on the system level. Therefore, it's vital to enroll the real data identification process for all proposed technics in order to make sure the potential adoption will not cause chaos.

Secondly, usually most of the parameters in a simulation are set according to some traditional statistical assumptions which are summarized from real measurements long time ago, such as the path loss models in wireless links and the Poisson distribution of user requests arrival. However, these kind of assumptions need to be verified or adjusted by new measurements from ongoing networks because of the rapid change of communication environments and human dynamics [86].

From this point of view, real data should play the role of standard in cellular networks analysis. Besides, unlike decades ago when large scale of measurements were very difficult to be conducted, nowadays data measuring and storing technologies are advanced enough to record every detail of the activities within the cellular networks . Therefore, real data is able to function as a standard to evaluate the various proposed techniques.

Real Data as a Solution

In fact, cellular networks are highly related to other human-related systems, like the transport system, the social networks and the city planning. Commonly, these kind of systems all stem from the daily dynamics of human beings thus showing more or less similarity with each other. From this point of view, real data measurements from cellular networks can be beneficial to solve strategic problems in those aforementioned fields.

In detail, the snapshot records of cellular networks provide a density map of mobile users within the coverage area, which is able to characterize the spatial distribution of corresponding transportation [12]. Therefore, combining the continuous snapshots description, the real data can depict the movement of mobile users, so are the related transport traffic which is really helpful for the transportation planning.

Also for social networks, cellular networks measurements is able to provide a holistic view of how the mobile users communicate with each other, thus tracing the link flows which supports the specific information gossip. Even more, based on timely real data, the monitoring system is able to detect the vital node which plays the key role in the link flows, and necessary steps can be implemented to cut or weaken the information gossip.

Considering the universality of cellular networks and the possible detailed records, real data measurement is able to provide an effective solution for many other human-related systems. In [79], the author built a clear correlation between the SINR level in cellular networks and the user engagement rate of video websites, giving specific suggestions for over-the-top (OTT) providers to improve the user experience in the network point of view. Return to the cellular networks themselves, real data is also kind of solution for the more and more important traffic prediction problem.

Actually, traffic prediction is crucial in cellular networks because its accuracy impacts the allocation of scarce communication resource [87], especially the frequency resource. Practically, historical traffic records in cellular networks can serve as the training set of the predict procedure, along with the increasingly powerful machine learning techniques, higher prediction accuracy can be achieved. Therefore, real data is promising to provide effective solutions for problems across from cellular network itself to various related human-oriented systems.

Conclusively, today's cellular networks are already very complicated in their implementations, not to mention the forthcoming 5G networks which are destined to involve multiple access technologies and satisfy multiple classes of service criterions. To simplify the designation of cellular networks and make them more efficient, data-driven approaches might be promising as real data can be adopted as a source, a standard and a solution for the evolution of cellular networks in our claim. After introducing the cellular network architecture and potential applications of real data, we will give a comprehensive review on the real traffic measurements in cellular networks. In this scope, three dimensions of traffic dynamics will be presented, i.e., time, space and content.

2.2 Real Traffic Measurements in Cellular Networks

After explaining the crucial importance of real data measurements in cellular networks, we will introduce the state-of-the-art research in this literature. Cellular networks traffic, along with the development of network infrastructures, are going through an evolution in dimensions of volume, type and demand patterns. As far as in 1997, author of [97] indicated the paradigm shift from traditional circuit-switched calls of Erlang-distributed patterns to advanced packet-switched multimedia services that appeals for new teletraffic model and service paradigm. Traffic in cellular networks, including message [106], voice [96] and data [83], has shifted from the voice-domination consuming pattern to the data-domination consuming pattern, resulting from the evolved user demand over years. Generally, the traffic as a wide-ranged variable, requires many features to characterize, from statistical description to geographical representation. In this section, to be clear and complete, we mainly cover three basic features, i.e., the temporal, spatial and content description of traffic in cellular networks. Actually, these three dimensions of characterization provide the most influential impact on the coordination between network function and user demand. As follows, we will introduce them one by one in specific details.

2.2.1 Temporal Characterization of Cellular Traffic

Traditionally, the traffic volume is considered to be homogeneous across different time scales for a cell, BTS or BSC, in order to simplify the planning procedure of cellular networks. However, the absolutely uniform assumption is definitely not true, thus light-load status are common for BSs since they are usually fuelled with the necessary radio resource to meet the highest traffic demand. On a smaller time scale, it is usually assumed that the call activities and data requests follow a Poisson process temporally, which means the arrival interval between subsequent calls is exponentially distributed. Actually, this kind of assumption can be true as A. K. Erlang claimed the Poisson distribution of telephone traffic one hundred years ago. However, as the

technology evolved from wired to wireless links, and cellular networks embraced multimedia as new approach for communications, the inherent patterns of traffic demand may also have changed. To identify the validation of traditional Poisson distribution or discover new suitable models, real data is the good and only way to work.

In temporal dimension, there are two different approaches to characterize the traffic dynamics in cellular networks, i.e., the macro and micro views. In detail, the macro view means to deal with the summation of traffic volume within a cell, a BS or larger coverage area, and of which the temporal variations on different time scales are analyzed. On the other hand, the micro view includes the traffic records on the levels of users, applications, requests or even packets as data source, and based on which specific performance metrics are investigated on a smaller time scale compared to the macro view. Besides, on the application of different approaches, analytic results from the macro view aim to provide guidance on the overall resource allocation, infrastructure deployment and other macroscopic visions of the cellular networks. While, differently, the specific examination of micro view tends to influence the internal parameter tuning or protocol design for refining the present flaws and improving the microscopic performance. In the following presentation, we will divide the related works into two different categories, according to the aforementioned views, and compare them with those correspondingly traditional assumptions.

Macro View

To analyze the traffic dynamics of cellular networks in a macro view, the aggregation traffic of different cells are the crucial variables, which can be added up under a coverage region of BSC or MSC. Here, we will introduce the related works according to the research timeline and corresponding logics.

Date back to 1999, Almeida et al. investigated the temporal variation of voice traffic from a GSM network in Lisbon [3], where they found a similar temporal pattern (highly related to the daily schedule of local residence) for different cells. Furthermore, they proposed to use the common double-gaussian and trapezoidal distribution to model the temporal variation of voice traffic, and showed that different models apply to different areas. To the best of our knowledge, this work was the first time to characterize the voice traffic of cellular networks in a temporal way, though the data or message traffic are not included.

Regarding the message, in [106], the authors performed a comprehensive measurement of the short messages service (SMS) in a national cellular networks in India, when the short messages were still popular in 2006. According to their results, 7.2% of the total messages sent by mobile users are requests to SMS services, and at least 10.1% of the total received messages are sent by content providers instead of mobile users, which is contradictory to our conventional assumption. Accordingly, based on the real measurements from telecommunication operator in

China, authors in [112] showed that the message traffic adopt a periodic pattern on the BSC level.

Different from the result in [3], authors in [96] found that the voice load of individual sectors varies significantly even within a few seconds in the worst case and there exists high variability of traffic volume even across sectors of the same cell. Therefore, the results tend to diverge in different coverage regions. Generally, the larger the investigated area, the more regular the aggregated traffic dynamic, due to the average effect. For example, [112] also depicted that the voice traffic within a BSC tends to be significantly periodic, therefore inducing a considerable predictability.

After 2007, as the worldwide spread of smart phones, more and more data traffic are generated, along with the decrease of traditional voice and message traffic. Therefore, the temporal analysis of traffic dynamics in cellular networks turns from voice duration to data consumption, which tends to exhibit a similar but more variable pattern.

Based on the data set spans one week in 2007 from a nation-wide network with thousands of BSs, authors in [68] showed that the aggregate network load exhibits a nice periodic behavior with relatively high loads during the day and the lowest load during midnight. On the contrary, individual BS loads do not show that much periodicity. Also, the load curve varies significantly among individual BSs with their peaks occurring at different times of the day. Similarly, this study on data traffic shows a regular pattern like the traditional voice traffic, along with the claim that the aggregation traffic in larger area tends to appear more periodic than individual smaller coverage area.

Utilizing a one-week-long data records in 2010 from a specific state in USA, in [83] authors found a diurnal characteristics of traffic volume over the duration of a complete week while weekdays tend to attract more user activity than weekend. Furthermore, the time-series of aggregate Internet traffic volume can be modeled by a multi-order discrete time Markov chain which would contribute to a good analytical property for performance evaluation.

Concluded from these above related studies, the aggregate traffic of large coverage areas in cellular networks tend to be significantly periodic which leads to high predictability, while the total traffic in smaller coverage areas (i.e. a BS) shows different levels of variation which could be attributed to human mobility and geographical diversity. To reach a thorough understanding into the traffic dynamics, more in-depth measurement should be conducted.

Such as in [93], Wang et al. presented a comprehensive analysis of the data traffic temporal dynamics across thousands of BSs from Shanghai, China. Using machine learning techniques, BSs can be clustered into five different categories based on their traffic dynamics on different time scales, where these categories are highly related to the geographical locations of cellular towers. Besides, they conducted a spectrum analysis on the frequency domain and claimed that

every traffic from the investigated BSs can be constructed using just four principal components corresponding to human activities.

Also in [92], the authors quantitatively characterized the spatio-temporal distribution of mobile traffic based on large-scale data set obtained from 380,000 BSs in Shanghai spanning over one month. They found that the mobile traffic loads uniformly follow a trimodal distribution, which is the combination of compound-exponential, power-law and exponential distributions, in terms of both spatial and temporal dimension with accuracy over 99%.

Furthermore, authors in [103] implemented a time series approach to analyze a large amount of traffic data from thousands of cellular towers, and they revealed that the mobile traffic temporal pattern can be divided into two components, regular and random parts. Based on the real data analysis, they discovered a high predictability of the regularity component of the traffic, and demonstrate that the prediction of randomness component of mobile traffic data is impossible.

Conclusively, the traffic dynamics of cellular networks on the macro view exhibit a diurnal pattern which can be adopted as a reference for the resource allocation on a large scale. Meanwhile, there still exists evident variation for respective traffic across different cells, which suggests that operators should conduct different resource allocation policies for different BSs, not like the one-for-all strategy. In the next subsection, we will introduce the traffic dynamics in a micro view.

Micro View

Different from the macro view for traffic dynamics analysis, the micro view provides a more delicate perspective to characterize the temporal properties, usually making use of user-level, application-level or even packet-level measurement records. Meanwhile, the corresponding traditional assumption in this literature is that the arrival pattern of traffic follows a Poisson process in the view of a BS or a specific content. As follows, we will uncover the realistic phenomenon discovered by many researchers in this field based on the measurements across the world.

Firstly, for the voice traffic, authors in [41] examined the call duration distribution in GSM networks. The result shows that the log-normal distribution is more precise than the exponential or Erlang distribution. As we know, the log-normal distribution preserves heavy-tailed property at some extent, which reflects the inhomogeneous nature of call durations in GSM networks.

Furthermore, authors in [96] certified this conclusion through another set of real data. In that study, they presented a large-scale characterization of primary users and found that the duration of calls are not exponential in nature and possess significant deviations that make them difficult to model. These two similar results based on different data set reveal the phenomenon which is clearly contrary to the traditional homogeneous assumption.

Secondly, we try to examine the similar property for data traffic. In [95], Williamson et al. used fine-grained measurements from an operational CDMA2000 (Code Division Multiple Access) cellular network to characterize the wireless Internet data traffic and revealed different patterns compared to traditional assumptions. In detail, the results show that the packet arrival process is non-stationary, and exhibits bursty nature rather than a homogeneous Poisson process. Related to this study, the authors in [108] compared the wireless traffic with traditional wired traffic, and found that the data sessions in wireless traffic contain less data in more but shorter flows, which typically consists of smaller packets with burstier arrival patterns. Similarly, these two studies revealed the same inhomogeneous property for data traffic as the voice traffic, both of which are high coupled with the human dynamics.

Go deep into the relationship between cellular networks traffic and human dynamics on temporal view, authors in [100] investigated the pattern of call activity during crowded soccer events in Brazil based on real measurements, and indicated that the dynamic transition of call activities between different cells can shed light on the inherent pattern of human mobility.

Also in [80], the author presented a first performance characterization of an operational cellular network during crowded events, where the temporal deviation of voice call and data traffic is compared to routine days. Besides the performance characterization during crowded events, the authors also proposed two effective methods to mitigate the severe performance degradation which are verified by simulation of real traces.

On the other hand, to link the personal traffic usage with mobile profiles, authors in [67] proposed a framework that automatically categorize mobile users into four different profiles according to their daily traffic usage, and based on which the authors calculated the traffic distribution of different kinds of users and create a traffic generator that captures the realistic usage pattern.

Conclusively, investigating the cellular traffic in the micro view provides insights about how the fine-grained traffic is requested and delivered in the networks. Contrary to the traditional exponential distribution for call durations and Poisson process for data traffic arrival, many recent investigations show inhomogeneous nature for cellular traffic on different dimensions.

Overall, the temporal analysis of cellular traffic based on real measurements in both macro and micro views challenges traditional assumptions on many important patterns, such as the equivalent assumptions for all BSs and the homogeneous assumptions for call duration and data arrival. According to these related works, the traffic consumption in cellular networks exhibits a periodic pattern for large coverage area which indicates significant predictability, while the traffic temporal pattern for single cells tends to be various across the cellular networks which urges different policy for distinctive BS. Meanwhile, the voice and data traffic in cellular networks exhibit inhomogeneous nature on different metrics, such as the heavy-tailed property for the

call duration and bursty inherence for data request, which would display significant importance in the network performance evaluation.

2.2.2 Spatial Characterization of Cellular Traffic

Besides the temporal variations, the spatial distribution of traffic consumption in cellular networks is far away from the uniform assumption, which is usually adopted in the network simulation of academic research. Actually, due to the mobility feature and aggregate residence of human beings, the mobile users are clusteringly distributed within the whole cellular networks. Thus, the traffic demand of mobile users inevitably varies across the overall coverage area. Specifically, the spatial distribution of traffic density is highly co-located with the human residential hot spots.

In practice, in order to manipulate the limited wireless resource efficiently, the operators need to pre-allocate the frequency bands and corresponding transmit powers to different BSs or cells. Obviously, the resource allocation strategy is necessary to be coincident with the real traffic demand, temporally and spatially. For example, the heavily loaded BSs desire more spectrum and transmit power, while the low-load BSs need less, and the specific quantitative allocation is related to the predicted traffic demand based on historical records. Furthermore, the inter-cell interference in cellular networks is also a challengeable issue, since it's highly related to the frequency reuse paradigm and transmit power allocation. Ideally, the frequency reuse factor should be various according to the realistic traffic load, because the interference level is dynamically changed. Therefore, the frequency reuse factor should be smaller in hotspot areas than that in low-load areas, and smaller at peak traffic period than that in low-load period in the same coverage area.

From these points of view, it's essential to distinguish the realistic spatial distribution of traffic demand from traditional assumptions based on real data measurements. Furthermore, using those empirical results for spatial distribution of traffic demand, it's possible to provide more efficient resource allocation strategies and interference mitigation methods to improve the overall capacity of cellular networks.

Actually, since a long time, researchers have found that the spatial distribution of traffic consumption in cellular networks was not uniform. For example, authors in [38] investigated the inhomogeneous property of voice traffic in GSM cellular networks, and proposed to use log-normal distribution to model the PDF (Probability Density Function) of traffic volume which can not be rejected in different levels of granularity. Besides, they also demonstrated that there is a distinctive capacity gap between homogeneous and inhomogeneous case, which highly motivated the accurate characterization of traffic distribution in order to provide reasonable guidance for cellular networks planning. This study investigated the numerical distribution

of traffic volume density regardless of corresponding correlation between human activities and traffic distribution.

In [3], besides the temporal dynamics, the authors discussed the spatial distribution of voice traffic in GSM networks, where they discovered the explicit inhomogeneous nature of traffic distribution and adopted several common functions to model this kind of nonuniform. From their results, farther is the cell from the city center, smaller is the traffic density in that area, which depicts a clear decaying phenomenon. Specifically, considering the relationship between the distance from the city center and the corresponding traffic density, the exponential model presents the worst performance while pairwise linear model outperforms the other candidates. Furthermore, authors in [91] proposed a demand-node generating model for infrastructure deployment in cellular networks, which considers the spatial aggregation effect of voice traffic across the coverage area which is highly correlated with human daily activities.

As cellular networks upgraded and the traffic usage of mobile users transit from the traditional voice calls to the popular data service, which stimulates the academic group to focus on the data traffic consumption across the whole network. For example, studies in [68] showed that 10% of the BSs experience roughly about 50-60% of the aggregate traffic load, which indicates the explicit imbalance of data traffic distribution across cellular networks. Besides, it also showed that less than 10% of subscribers generate 90% of the load which reflects the clustering effect of traffic distribution on another dimension. This phenomenon from real measurements clearly demonstrated the heavy-tailed characteristics of data traffic on the BS-level and user-level, which is inherently rooted in human dynamics (similar with Matthew effect in economics and sociology).

More specifically, Laner et al. in [48] investigated the real data from a high-speed packet access (HSPA) networks in Vienna (Austria), and found that the mean throughput of coverage cells within the peak hour varies over roughly three orders of magnitude. Whereas the 10% of cells with lowest load have a mean throughput of below 1 kbit/s, the 10% most loaded cells have a mean throughput above 500 kbps. Consequently, the commom assumption of a constant traffic density over a large number of cells is inadequate.

Besides the aggregate traffic, Shafiq et al. in [82] analyzed the geospatial dynamics of application usages in a large 3G cellular networks, based on traces from both RAN and CN indicating location information and data delivery details. Based on the application usage calculation on different levels, the authors aimed to classify a number of BSs into different categories. Counter-intuitively, they found that cell clustering results were significantly different for traffic volume in terms of byte, packet, flow count and the popularity of different applications significantly varies even within a given neighborhood.

From the intangible heavy-tailed description to detailed statistical modeling, Lee et. al in [50] found that the cell traffic can be approximated by the Weibull or Gamma distribution; the traffic density can be approximated by the log-normal and Weibull distribution which are all contradicted with the traditional uniform assumption. Besides, they also found that there is a correlation between the traffic volume of different cells within specific distance. More importantly, this study provides the possibility of generating the realistic traffic demand across the whole plane of cellular networks, embedded with the traffic correlation between BSs and without losing the inhomogeneity characteristics.

To consider the mobile user distribution and spatial traffic demand jointly, many researchers started to use fine-grained user distribution information to shed light on the inherent relationship between user and traffic. Ding et al. in [28] discovered that the spatial distribution of subscribers and traffic demand can be accurately described by log-normal mixture models. Besides, their extensive analysis gave a precise characterization of BS capacities and clustered all BSs into six categories based on subscriber density and average traffic demand. These kind of results may help the operators to allocate their limited resources more efficiently based on the traffic demand categories of different BSs.

On the other hand, to combine the traffic consumption with the mobile user distribution and the infrastructure deployment, authors in [62] proposed a tunable statistical model capturing the interconnection between BS spatial deployment and spatial traffic distribution, with only two parameters. This work provides a convenient way to connect a heterogeneous infrastructure deployment and the corresponding spatially heterogeneous distributed traffic demand.

Conclusively, similarly with the inhomogeneous distribution of traffic demand on temporal dimension, the aggregate traffic consumption presents heterogeneity across the spatial plane within the cellular networks. Specifically, the time-summation traffic volumes of different BSs exhibit heavy-tailed characteristics according to numerous real measurements verification. This phenomenon indicates the severe imbalance of traffic demand across the whole networks, thus urges the load balance technics or different resource allocation strategies to improve the overall capacity performance.

Furthermore, compared with the aggregate traffic volume on BS level, the traffic density description provides a more intuitive view into the heterogeneous nature of traffic consumption. Based on different data sets from different countries, the state-of-the-art statistical model for the data traffic density in cellular networks follows a log-normal or a Weibull distribution, both of which shows some extent of heavy-tailed property thus verifies the spatial heterogeneity nature of traffic demand more definitely.

Actually, analyzing the spatial distribution of traffic volume on the one-dimension numerical statistic is not enough since it loses explicit location information where the traffic demand

actually happens. For example, despite the log-normal distribution may be accurate for traffic density values, but it's not adequate for locating the data requests or the aggregate traffic on user or BS-level. One possible way to solve this problem is to spatially modeling the user distribution or BS deployment on a two-dimension view, and then feed those information into the traffic density description. In this point of view, the spatial heterogeneities of mobile users, BSs and traffic demand are coupled with each other and should be investigated all together which leads to our works on spatial modeling of BSs in Chapter 3.

2.2.3 Content Preference of Cellular Traffic

Traditionally, before the 3G cellular networks, the available bandwidth was not sufficient for mobile users to fetch contents directly from the Internet. At that time, the main goal of wireless communications was mainly the voice call or short message services between subscribers. Meanwhile, the wired network was speeding up to provide various content options through the Internet, such as news, pictures and videos. Looking backward, during the past decades, the provided contents and the connected Internet help each other to spread and enrich, and their combination makes our real life and virtual activities seamlessly merged together. The social pattern of us, as human beings in the Internet era, reflects itself from the touchable real life onto the virtual binary world. Therefore, the wired Internet or mobile wireless network can be utilized as mirrors to reflect the inherent human dynamics, thanks to their well organized recording ands almost ubiquitous coverage.

Firstly, for the wired caching, Lee et al. in [11] first showed that the web requests follow a Zipf-like distribution [65]. Based on various traces collected independently, the authors introduced a simple model for web requests, which are independently valued from a Zipf-like distribution. The results showed that this simple model can explain the asymptotic behavior of these three properties that are observed in real web cache traces.

Apart from the traditional web content, the authors in [36] conducted an extensive analysis of the YouTube workload, and found that there are (not surprisingly) many similarities to traditional web and media streaming workloads. For example, since access patterns are strongly correlated to human behaviors, as traffic volumes vary significantly by time-of-day, day-of-week, as well as longer term activities (e.g., academic calendars). Similarly, video files are much larger than other types of file, and some videos are more popular than others.

Similarly, Cha et al. in [14] presented an extensive data-driven analysis on the popularity distribution, popularity evolution, and content duplication of user-generated video contents on the Internet. They studied the nature of the user behavior and identified the key elements that shape the popularity distribution, and it was found that the Pareto phenomenon of content requesting is likely caused by both human similarity and the information filtering technics.

Summing up the related works on the web content across the last decade, no matter what kind of them, the contents popularity distribution exhibit similar aggregated feature (heavy-tail distribution), although there are some differences between the requesting pattern.

More comprehensively, [81] presented a measurement study of a large commercial content delivery networks (CDN) serving thousands of content providers, and they found that the top 1% content publishers account for up to 60% of the total request counts for both small and large object platforms. Moreover, top 1% content publishers account for more than 90% of the total request size for the large object platform, while more than 60% and 50% objects in the large and small object platforms, respectively, are requested only once.

Conclusively, the content requests on the Internet have a significant tendency to be aggregated on different scales, such as the content popularity, content provider and content duplication. For the mathematical characterization of the popularity distribution, the Zipf-like models present the best performance.

Although the access technology is far different between the broadband wired Internet and mobile wireless networks, the content requests exhibit similar aggregated properties, specifically after the speed-up of cellular networks. On one hand, for the categories of requested contents, the mobile users are able to access most of the contents on the Internet, including web pages, audio, video and so on. On the other hand, for each category or in the global view, the popularity of different contents also presents an unbalanced phenomenon.

Such as in [83], based on a large amount of real data from cellular networks, the authors discovered that the distribution of network traffic with respect to both individual devices and constituent applications is highly skewed. Only 5% of the devices are responsible for 90% of the total network traffic. Moreover, the top 10% applications account for more than 99% of the flows. More specifically, these distributions for the popularity of different ranked contents can be modeled using Zipf-like models, which is the same as in wired networks.

Similarly in [46], Lin et al. utilized one month's trace data from a US operator, and characterized the data usage pattern in a large UMTS cellular networks. In accordance with the expectation, they found that a few users (top 3%) consume nearly half of the total data traffic, and exhibit distinctive usage patterns compared to normal users. Furthermore, among these heavy users' usages, a small number of dominant applications make up the majority of data traffic, including mobile video/audio sites, social networks and popular mobile applications.

Authors in [31] examined the video traffic generated by 3 million users in 2011 across one of the largest 3G cellular networks in US, and they found that video traffic accounts for 30% of the downstream cellular traffic during busy hours. Besides, the results also showed that 77% of the traffic is concentrated in just the top 10 content providers and 24% of the bytes for progressive downloads requests can be served from cache. Conclusively, the data traffic in cellular networks

exhibits significant clustering feature not only on the content categories but also on the user usage level, and the popularity of different contents can be modeled as Zipf's law.

Besides the normal time, the clustering nature of content requests expresses more aggressive during the busy hour. For example, in [32], the authors investigated the traffic dynamics during the 2013 Super Bowl event in New Orleans based on detailed records from the overlapped LTE networks. From their study, the most popular applications during this tremendous event are web browsing and video streaming, and the most accessed content providers are cloud providers which account for over 22.1% of the overall traffic. Even more dominant than that in broadband networks, the HyperText transfer protocol (HTTP) traffic constitutes more than 90% of the total multimedia traffic in cellular networks, as depicted in [31] and [58]. From these observations, we can see that the contents preference phenomenon is tied up with the request procedure, especially on the base of a large number of mobile users, no matter in busy hour or normal usage.

Furthermore, after revealing the content preference of mobile users in cellular networks, it's also important to investigate the predictability of different kinds of contents and their performance impact on network utilization. Authors in [112] collected large amount of real data from a GSM/UMTS hybrid cellular networks, and investigated the predictability of three kinds of traffic, i.e., message, voice and data. According to their results, voice traffic has the highest predictability among these three, while nearby BSs and historical records can improve the traffic predictability significantly. Besides, in [108], the authors examined a wide range of services in cellular network traffic, and found that different applications impose different demands on network resources on packet level, flow level and session level.

Conclusively, the user requests in cellular networks tend to exhibit content preferences, showing explicit heavy-tailed distribution on content popularity. Together with the spatial and temporal dimensions, the dimension of content forms the complete clustering nature of traffic distribution in cellular networks. After introducing the separate characterization of each dimension, we will investigate the traffic dynamics on any combinations of these three dimensions, and shed light on the relationship between the clusterings on different dimensions and their potential application on the service capacity optimization in cellular networks.

2.3 Joint Characterization of Different Dimensions

After introducing the real measurements of traffic demand in cellular networks in three different dimensions separately, in this section we try to extend the analysis to combinations of different dimensions. Other than theoretical characterization, description of the combined clustering nature on different dimensions can lead to potential technic solutions.

2.3.1 Spatial-Temporal Characterization

In cellular networks, clustering properties are not only exhibited in the spatial domain, but also on temporal dimension, as illustrated in previous sections. Whether there is a dependence between these two dimensions is crucial for combining them together for more efficient service policies. Hereafter, we try to uncover the possible correlation between spatial distribution and temporal distribution of traffic demand in cellular networks.

Firstly, we need to figure out what is spatial-temporal clustering. In part of it, spatial clustering means that the traffic volume varies in different locations, and small portion of BSs occupy most of the overall traffic. For the second half, temporal clustering means that the traffic demand varies during different time, implying peak and valley, even though exhibiting some extent of predictability [112]. Combining them together, the spatial dimension of traffic demand may exhibit distinctive degrees of clustering effect at different time periods, such as office time and dinner time. On the other hand, the temporal dimension of traffic demand may exhibit distinctive degrees of clustering effect at different locations, such as residence and offices. Therefore, it can be declared that spatial dimension is coupled with temporal dimension showing an adjoint clustering phenomenon.

Secondly, how to deal with the correlation between spatial and temporal clustering? Generally, it's a mathematical problem dealing with two random variables. Therefore, we need to separately define the degree of spatial clustering and temporal clustering in a mathematical way, from which we can obtain two series of random variables. Then, we can conduct the formal correlated coefficient calculation for these two series, and obtain a number indicating the degree of correlation between them.

Finally, how to make use of this correlated phenomenon? Intuitively, in cellular networks, we can switch off some BSs in low traffic region when the spatial clustering effect is quite obvious, while ensuring the coverage function through nearby BSs. However, during this operation, it's necessary that the overall traffic volume is low in order to make sure that nearby BSs are capable of delivering additional capacity. That's to say, the coordination between spatial and temporal is essential for BS-switch operation in cellular networks.

Conclusively, spatial clustering of traffic demand is coupled with that of temporal dimension, and this feature is promising for more energy efficient operations in cellular networks.

2.3.2 Spatial and Content Combination

Besides space and time records, content is another essential dimension for traffic demand, and it's getting more and more important regarding the increasing research of content centric networking. Originally, it's not welcomed that users tend to request the same content, because

it increases the average delivery time according to queuing theory, especially for client-server paradigm. After that, due to development of distributed sharing system and CDN technology, it shows as a good sign for service management, since the more concentrated are the content requests, the more efficient the delivery process. However, when it comes to the combination of content preference and spatial clustering, what is the effect on service capacity?

Firstly, like the case for spatial and temporal clustering, we need to figure out what is the meaning of clustering on both dimensions separately. As addressed, spatial clustering means the imbalance on the traffic distribution at different locations, and content clustering means that most of the user requests are directed to small amount of contents. After that, is there any correlation between spatial clustering and content preference? Or to say, for different locations, is the corresponding popular content also differs from each other? This claim needs to be verified by real measurements.

Secondly, how to deal with the combination of spatial and content dimension? If the content preference varies across the cellular networks, then how to characterize it mathematically? As we know, the content preference are characterized by the Zipf's law separately, and the degree of clustering is then indicated by the exponent of the popularity distribution γ. Here, after introducing the spatial disparity, we then obtain a series of γ. From this series, we can get a parameter representing the variation of content preference across the whole networks.

Finally, how to get benefits through compensation between content and spatial clustering? For example, the caching technologies are popular research topics in cellular networks in recent years. In RAN caching, BSs are assumed to cache limited amount of contents in their local storage, and the contents can be directly delivered to mobile users if requested. In this scenario, spatial clustering of BS deployment may be beneficial since the nearby BSs can perform as a cluster in order to enlarge the caching storage, thus storing more popular contents.

2.3.3 Temporal and Content Combination

As described in previous subsection, the clustering preference of traffic demand varies across the cellular networks, and it also changes on the time dimension. Therefore, we need to consider them together in order to track the variation of content preference in a time series way.

Firstly, the degree of content preference in cellular networks are well characterized by the exponent of Zipf's law (i.e., γ). On the other hand, there is plenty of traditional ways to analyze time series consisted by the Zipf's exponents. However, there is some information lost here, since the exponent is just dealing with the popularity proportion, without showing the order of exact popular contents. Therefore, in this thesis we assume that the content library and the absolute order of contents are all fixed, and the popularity distribution is our main concern.

Secondly, how to combine the analysis of temporal clustering and content preference together? Not only content popularity are various along the time dimension, but also the traffic summation of content requests are dynamic. Therefore, consider these two effects together, the traffic demand for each content would be tremendously various on the time scale. Similarly with the combination of spatial and content clustering, it's possible to use the joint probability to characterize the merge of temporal clustering and content preference.

Finally, how to make use of the combination of temporal clustering and content preference? As we know, the static content preference properties are beneficial for the traffic management in distributed systems, where caching technology can play a crucial role. After combining on time dimension, if the traffic demand of a specific content achieves a predefined threshold, we can adopt multicast to perform 'one transmit multiple receive" paradigm. Due to the openness nature of wireless communications, the broadcasting technics could be beneficial for improving the service capacity.

2.4 Conclusion

After introducing the real measurement works in different literatures, we can conclude that the clustering nature is widespread in cellular networks, spanning from user traffic to infrastructure deployment. In this chapter, we carefully examined the statistical features of traffic demand, from temporal characterization to spatial distribution, and then the content preference description. For different dimensions, we presented plenty of related works which make use of real data from all over the world, thus demonstrating a comprehensive picture on the realistic scenarios in cellular networks. Accordingly, we divided the conclusion for the measurement results into three parts, namely temporal, spatial and content preference.

Firstly, the temporal analysis of cellular traffic based on real measurements in both macro and micro views challenges our traditional assumptions on many scenarios, such as the equivalent assumption for all BSs and the homogeneous assumptions for call duration and data arrivals. Specifically, the traffic consumption in cellular networks exhibits a periodic pattern for large coverage area which indicates significant predictability, while the temporal traffic pattern for single cells tends to be various across cellular networks which urges different resource policy for distinctive BS. Meanwhile, the voice and data traffic in cellular networks exhibit inhomogeneous nature on different metrics, such as the heavy-tailed property for the call duration and bursty inherence for data request, which would display significant importance in the network performance evaluation.

Secondly, similarly to the inhomogeneous distribution of traffic demand on temporal dimension, the aggregate traffic consumption presents heterogeneity across the spatial plane within

the cellular networks. Specifically, the time-summation traffic volumes of different BSs exhibit heavy-tailed characteristics according to numerous real measurement verification. This phenomenon indicates the severe imbalance of traffic demand across the whole networks, thus urges the load balance technics or differential resource allocation strategies to improve the overall capacity performance.

Thirdly, the user requests in cellular networks tend to exhibit content preferences, showing explicit heavy-tailed phenomenon on content popularity which can be accurately modeled by Zipf distribution. Together with the spatial and temporal dimensions, the dimension of content forms the complete clustering nature of traffic distribution in cellular networks. After introducing the separate characterization of each dimension, we investigated the traffic dynamics on any combination of these three dimensions, and shed light on the relationship between the clusterings on the different dimensions and their potential application on the service capacity improvement in cellular networks.

Finally, by combining these different dimensions of clustering properties together, it's possible to make better use of the limited resources in cellular networks. In the following chapters, after examining the clustering nature more specifically and describe them in a mathematical way, we try to introduce promising technics to improve the efficiency of the whole system.

Chapter 3

Spatial Modeling of Base Stations: Challenging the Convention

Contents

Cellular networks provide services through the wireless link between the BSs and mobile users, which is the most fundamental and challengeable part of the whole system. In the radio access networks, limited communication resources like frequency and power, are partitioned to different

BSs and mobile users partly based on the channel state and interference level which is highly correlated to the locations of BSs and users. That's to say, the spatial properties of cellular networks play an essential role in the wireless communication service.

The following contents will be divided into six sections, where the first one will introduce the motivation and overview of related works, then the detailed real data description will be presented in Section 3.2. After that, we will introduce the point process models and the evaluation metrics in Section 3.3, followed by the fitting procedure description and accuracy examination in Section 3.4. In detail, in the modeling processes across this chapter, we investigate both the urban and rural districts, the macro and micro BSs, small regions and large areas, the singular data sample and multiple large-scale samples as shown in Section 3.5. After all these, we conclude this chapter and indicate the uncovered problems, which leads to the works in next chapter.

3.1 Introduction

As said, the spatial structure of BSs has a great impact on the performance of cellular networks, since the received signal strength varies depending on the distance between transmitter and receiver [42]. Moreover, interference characterization is very complicated and challenging due to the path loss and multipath fading effect, in particular for a heterogeneous networking scenario consisting of different types of BSs. In order to evaluate the network performance more accurately, it is essential to obtain realistic spatial models for BSs deployment in cellular networks [5].

3.1.1 Background

Beside the multipath phenomenon in wireless transmission of cellular networks, the distance-based path loss effect contributes to the signal propagation most significantly. The deployed BS can only serve a range of mobile users within its coverage area, due to the signal strength decreases with the distance between user and BS. Therefore, the BS locations play a key role in the coverage performance of cellular networks, since it determines the signal strength and thus the capacity of deployed cell. In this term, the spatial distribution of BSs is a key driver for cellular network capacity enhancement.

Besides, as the traffic demand increases exponentially, the cellular networks are transforming from uniform deployment to heterogeneous layout, i.e., from coverage-driven macrocells to capacity-driven small cells. Although this densification method can improve the throughput of the overall network by increasing the frequency reuse factor, it inevitably enhances the interference level of each connection since there are more interferers nearby.

In terms of necessity and complexity, it's important to conduct a comprehensive experiment on the spatial distribution of BSs in cellular networks, using a large amount of real data and realistic spatial point process models. Before presenting our approaches, we would like to introduce the related works which helps us to understand the difficulties and trade-offs in this problem.

3.1.2 Related Works

By far, hexagonal grid model has been popular to model BS locations in academia and industry due to its simplicity and regularity [71]. Although simple, this model captures several key aspects of cellular networks and has been an industry standard. However, despite its simplicity, it is surprisingly intractable, especially for the downlink analysis, and thus is used mainly for the system-level simulations [17]. Furthermore, as cellular networks have evolved for decades, the real BSs deployment is significantly influenced by population and landform, which makes the regular grid assumption even more impractical.

To solve this problem, in recent years, PPP* has been proposed to model various network structures [4, 5, 27, 42]. Different from the deterministic grid models, PPP characterizes the BS locations in a stochastic way. That's to say, it doesn't specify the exact position for each BS, but provide a spatial realization for each sample area. Meanwhile, the number of BSs in PPP is not constant either, but comes from a Poisson distribution whose mean indicates the density of BSs in that region. As a baseline role, PPP model can provide tractable and useful results for performance evaluation in both one-tier and multi-tier networking scenarios [15, 18]. Furthermore, it helps to derive the close-form performance characterization of many technics proposed in RAN [44, 105]. However, it may not be the most suitable one to model BS locations as researchers hardly reach a consensus on PPP's performance to model the real deployment. For example in [4] and [49], the authors observed inconsistent coverage probability performance of the PPP models for real BS locations from different cities around the world. Given these conflicting results above, it is still worthwhile to conduct more comprehensive investigations to provide convincible conclusions, and take more realistic models into consideration.

Generally, in stochastic geometry literature, despite of PPP's mathematical perfection, there are plenty of choices including repulsive and clustered point processes [20] to model various spatial patterns. Repulsive point process avoids the included nodes to be too close from each other, and clustered point process put the nodes together thus leaves an inhomogeneous appearance. For example, in [73], the authors discovered that the Geyer saturation process, which takes account of pairwise interaction between points, can accurately reproduce the spatial structure of various wireless networks. More specifically in cellular networks, Geyer saturation process and its special case Strauss process are utilized to model macrocellular deployment for different

*Poisson point process.

scenarios in [89]. Besides, Poisson hard-core process (PHCP) is also proposed in [40] and Poisson cluster process is verified to be able to model BSs deployment in urban areas [49]. Similarly, the Ginibre point processes and determinantal point processes have been investigated as suitable models for wireless networks with nodes repulsion [25, 56], obtaining a tentative compromise between accuracy and tractability. In summary, various point processes have been employed to model BSs spatial structure based on different data sets from cellular networks worldwide [29], but the conclusion is still indistinct so far in this literature, due to the considerable insufficiency of real data samples and the significant disparity between different cellular networks.

Indeed, the spatial distribution of BSs in cellular networks is far more complicated than what is commonly expected. Firstly, various regions such as rural and urban areas are deemed as distinctively different cases, owing to population density divergence and disparate traffic demands [89]. Secondly, because of the limitation on site selection, the human factor and geographical effect have significant impacts on BSs spatial distribution. Thirdly, for heterogeneous multi-tier cellular networks, BSs on each tier differs in transmit power and coverage area thus the spatial distribution varies for different tiers [21, 26].

Specifically, as cellular networks undergoing an evolution towards heterogeneous networking architecture [18, 27], considering the inherent differences on functionality and networking feature, the spatial structures of macrocell and microcell may have different layouts. Actually, macrocell BSs are neither too close nor too far away from each other in order to satisfy coverage requirement and decrease inter-cell interference. Therefore, there is a repulsion between macrocell BSs. On the other hand, microcells are usually deployed to diminish coverage hole and offload network traffic, which always exhibits aggregation feature. So microcell BSs would be clustered. These facts provide reasonable basis to adopt Gibbs and Neyman-Scott processes [20] to model macrocell and microcell deployment, respectively.

Besides the functional difference between BSs, the geographical factor also gives rise to the disparity of BS spatial distribution. For example, BSs' distribution in urban areas are much denser than that of rural areas, and BSs' density in mountain or river areas are much less than that in flat land.

3.1.3 Our Approaches and Contributions

Our Approaches

In order to solve these challenging problems, massive real data is essential to provide a holistic view on the spatial distribution of BSs. Moreover, due to its complexity, a reasonable proposition may be different point processes work for different scenarios, such as repulsive models for macrocells and cluster models for microcells [99, 115, 116].

Actually, there are three steps to complete the realistic spatial modeling process, i.e., the data collecting, the model selection and the fitting procedure. After collecting massive real data from on-operating cellular networks, we need to choose suitable point process models for these two-dimension data sets, based on preliminary examination on the data samples. Then, the chosen models are fitted to the real data, and specific parameters and performance metrics can be derived for the final accuracy evaluation.

In addition to the real data sets used in BSs spatial characterization, the statistical modeling process itself entails a two-fold preparation. The first component is the point processes selected to be fitted to real data, the other one is the performance evaluation metrics utilized for model selection. Actually, in some cases, BSs are neither too close nor too far away so as to guarantee full coverage and mitigate inter-cell interference. This phenomenon makes it reasonable to utilize Gibbs point processes, which can describe the repulsive property. Besides, in some dense urban areas, BSs tend to be aggregately distributed in order to provide high capacity requirement for more subscribers, thus Neyman-Scott point processes are chosen to model these clustering nature. Therefore, these three kinds of point processes [20] are adopted as candidate models besides PPP in this work.

Moreover, two types of metrics categorization, namely statistical metrics and network-layer performance metrics, are adopted for hypothesis testing. The widely applied statistical metric is Ripley's K-function or its transformation L-function [74], while the coverage probability is the most popular metric of performance evaluation due to its fundamental usage in wireless network analysis. Details can be found in following section of evaluation metrics.

Given these spatial model candidates and evaluation metrics, we provide the spatial modeling of all the BSs from different perspectives. Specifically, we divide the overall BSs dataset into disjoint subsets according to geographical factors (e.g. rural or urban areas) and functional types (macrocells and microcells), respectively.

Our Contributions

The objective of this chapter is to provide realistic spatial models for BS locations in cellular networks. Compared to the literature, the advantages in our approach are three-fold. Firstly, our work is based on massive real BSs deployment data from the largest telecommunications operator in China, and thousands of regions are randomly selected to identify different point processes. The huge amount of data source ensures the accuracy and universality of the resulting models. Secondly, typical models including PPP, Gibbs point processes and Neyman-Scott point processes are adopted as candidates which are compared in terms of two types of performance metrics. Thirdly, separate modeling processes are conducted for different tiers and different

regions within the heterogeneous cellular networks. Accordingly, our technical contributions in this chapter are three-fold as well.

- The accuracy of PPP for BS locations in cellular networks is questioned by our extensive identification. This result will strongly challenge the popular adoption of PPP model in literature.
- The general clustering nature of BS locations is revealed by random verifications, and the degree of clustering varies significantly for BSs in different areas or different tiers.
- In general, it's verified that Neyman-Scott point processes have superior modeling accuracy than Gibbs point processes, while all of them are not sufficient for realistic characterization.

Following, we will introduce the real data set first, including small sample regions and large coverage areas and their corresponding population and geography information.

3.2 Real Data Description

In order to obtain an accurate and realistic model for BSs deployment, our work is based on a massive amount of real data including all BS-related records from the largest cellular networks operator. The corresponding province in China has a population up to 54.77 million with a 526 persons per square kilometer density. Within this 104,141 square kilometers province, the data set includes 47663 BSs of GSM cellular networks with more than 40 million subscribers, and each record of the BS contains the corresponding coverage area, location information (i.e., longitude, latitude) and type information (i.e., macrocell or microcell).

Based on the coverage area and location information, we can divide the dataset into disjoint subsets. For example, we obtain the subsets of urban areas and rural areas, by matching the BS information with local maps. In this chapter, for representativeness and integrality, we mainly consider three typical urban areas and one large rural area to examine the accuracy of various candidate models for BS locations. The population of these selected urban areas are three-layered, ranging from 1 million to 5 million, covering the so-called metropolis city (city A), big city (city B) and medium city (city C). Two of them (city B, C) are coastal cities, while the other one (city A) is inland city. Besides, the rural area covers a large portion of the central part of this province. The detailed information of these selected areas are summarized in Table 3.1.

From Table 3.1, we can observe that the BSs deployed in urban areas are much more denser than those of the rural area, so does the percentage of microcells in all BSs. As follows, we will introduce these sample regions in detail.

TABLE 3.1: Information of Selected Large Regions.

Region	Area (km^2)	BS number	Macrocell	Microcell	BS density (km^{-2})
City A	60×40	6251	3513	2738	2.604
City B	40×40	977	677	300	0.611
City C	30×50	1911	1538	373	1.274
Rural	200×200	12691	11603	1088	0.317

3.2.1 BS Locations in Large-scale Areas

As we proposed, the modeling procedure of this work will be conducted in different dimensions, from urban area to rural area and from macrocells to microcells. Besides, the experiment will also include specific sample examination and large-scale identification, which takes small sample regions and large sample areas into consideration, respectively. Therefore, in this subsection, we will firstly present the BS locations in large-scale areas.

BSs in Urban Area

As displayed in Table 3.1, there are 3 big cities in the whole data set. The areas, the number of macrocells and microcells, and corresponding densities are all given. Here, we take part of city A as an example to show the spatial distribution of BSs in urban areas, as depicted in Fig. 3.1.

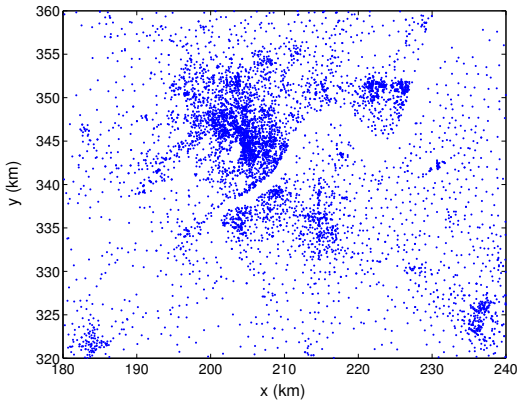


FIGURE 3.1: BS locations in city A.

Actually, city A is the capital of our investigated province with area 16,596 square kilometers, and its population is more than 9 million. As we can see in the landscape, city A contains dense urban areas and suburban areas, which can be easily noticed from the BS density variation across the figure. In Fig. 3.1, we sample a 60×40 rectangular region (2400 square kilometers) to represent the BS distribution in city A. This sampled region contains 6251 BSs in total, where 3513 of them are macrocells with the left 2738 ones microcells, and the density reaches 2.604 per square kilometers. Besides, city B and C are also large-scale urban area samples in our analysis, and all of them will be investigated in the following identification process.

BSs in Rural Area

Like urban areas, real data from rural areas are also essential for the realistic modeling of BS locations. In Fig. 3.2, we depict a large sample region from the rural area, which is mainly located in the central part of the investigated province. The rural region we selected are 200×200 square kilometers, which is much broader than the urban ones. In this region, the total number of BSs is 12691, and most of them are macrocells (11603).

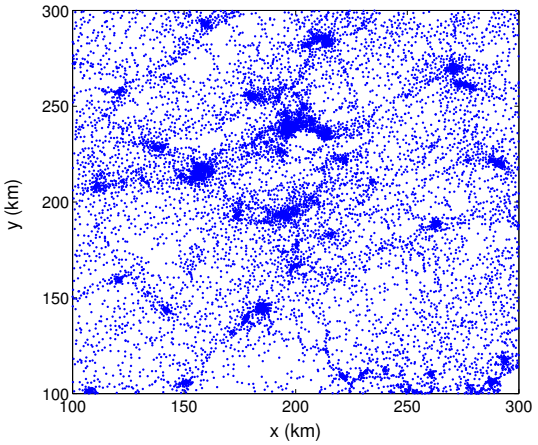


FIGURE 3.2: BS locations in the large rural area.

By combining urban areas and rural areas in the modeling process together, we can get to the difference of BS spatial distribution for areas with distinctive population and geography.

3.2.2 Small-region Samples of BS Locations

Besides large areas, small-region samples provide another description of BS locations on a different scale. Here, we sampled two regions to reveal the spatial distribution , where one of them is from the urban area (city A), and the other one is from the rural area.

Sample Region in Urban Area

As we know, the BS distribution in urban areas is much denser than that in rural areas. Therefore, we decide to sample a small region from urban area, namely 3×3 km^2 from city A, as depicted in Fig. 3.3. This selected urban region contains 249 BSs including 84 macrocells and 165 microcells, while the high percentage of microcells reflects the great capacity demand in this dense urban region.

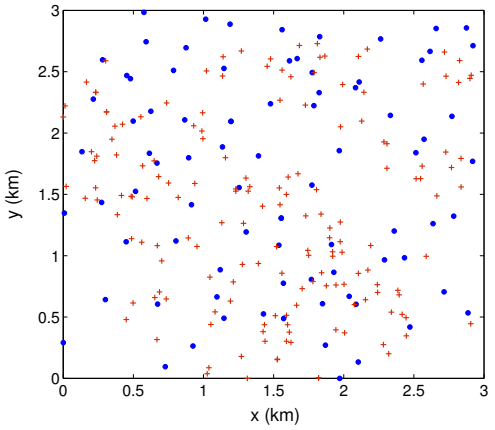


FIGURE 3.3: BS locations in a chosen dense urban region of city A, the blue dot represents the macro BS while red cross is the micro BS.

Sample Region in Rural Area

Unlike the sample region from urban area, we select a much broader region in rural area. Since the BS density is relatively smaller, and the number of BSs needs to be large enough to conduct the modeling process, we choose a 20×20 km^2 region as a rural sample. As depicted in Fig. 3.4, this selected rural area contains 79 BSs with only 5 microcells. The low density of BSs and even fewer microcells in rural area express the relatively higher demand for network coverage than capacity enhancement.

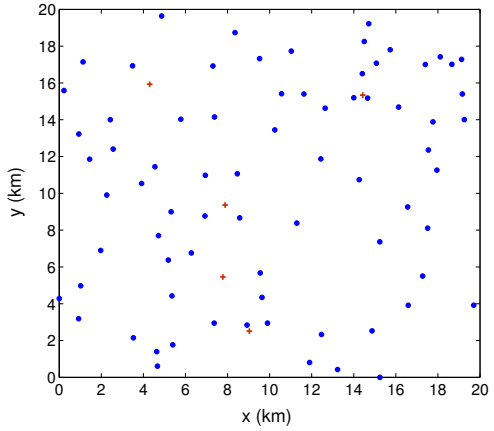


FIGURE 3.4: BS locations in a chosen rural region from a central inland part of the province, the blue dot represents the macro BS while red cross is the micro BS.

After presenting the real data sets for our modeling process, next step is to introduce the point process models and the evaluation metrics for the fitting performance. In the following sections, we will first give a brief introduction of point process models in stochastic geometry, including their difference and similarities. After that, we will introduce several common evaluation

metrics for our fitting procedure, including traditional statistical metrics and cellular networks performance metrics.

3.3 Spatial Point Process Models

The most basic component in stochastic geometry are the spatial point processes, which result in different network topologies. Intuitively, point process is a collection of points distributed in a selected window on the plane. More formally, it can be interpreted as a measurable mapping from a certain probability space to the space of point measures. In general cases, the point process can be represented as a countable random set $\Phi = \{z_1, z_2, ...\}$, of which the intensity measure Λ of Φ is defined as $\Lambda(B) = \mathbb{E}^{\dagger}\{\Phi(B)\}$, where B is a sub region of Φ and $\Phi(B)$ denotes the number of points in B. There are many kinds of point processes, such as the PPP, Hard-core processes, Gibbs processes, Neyman-Scott processes and the Cox processes [20, 66]. They can also be categorized into three sets, the completely random processes, regular processes and clustered processes. Among the regular point processes where repulsion is exhibited, Gibbs processes take a large part of them. Neyman-Scott process is a very typical class in clustered point processes, where there is attraction between points. Since real BSs deployment may be regular or clustered across the networks, and different regions may have different distribution patterns, we consider all kinds of models in this chapter to find the most suitable ones. To be more clear, a tree structure of these different point process models are given in Fig. 3.5.

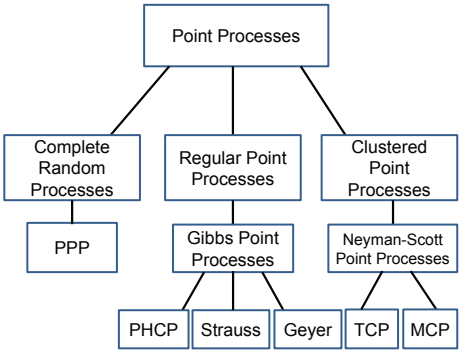


FIGURE 3.5: Tree structure of different point process models.

3.3.1 Completely Random Processes

Poisson point process is a complete random point process where no repulsion or attraction is depicted between any points.

[†]Expection of random variable.

Poisson Point Processes

Let Λ be a locally finite measure on some metric space $\mathbf{E}$, a point process Φ is Poisson on $\mathbf{E}$ if:

(1) For every bounded closed set B, $\Phi(B)$ follows a Poisson distribution with mean $\lambda|B|$, where λ is the density of this point process and $|B|$ is the area of B.

(2) For disjoint closed subsets B_1, B_2,...,B_n, the numbers of points in each subset $\Phi(B_1)$, $\Phi(B_2)$,..., $\Phi(B_n)$ are independent.

3.3.2 Regular Point Processes

Among regular point processes where there is repulsion between nearby nodes, Gibbs point processes are important branches in the stochastic geometry literature [20]. They are also referred as Markov point processes, because their property can be characterized by probability density, which is helpful in fitting and simulation using Monte Carlo method. Without loss of generality, we consider a point pattern $\mathbf{z} = \{z_1, z_2, ..., z_{n(\mathbf{z})}\}$ placed in a bounded window W, where $n(\mathbf{z})$ is the number of points in $\mathbf{z}$. For simplicity, only pairwise interaction is considered here, and its probability density function (PDF) can be defined as:

$$f(\mathbf{z}) = \alpha \cdot [\prod_{i=1}^{n(\mathbf{z})} \mu(z_i)] \cdot [\prod_{i<j} \rho(z_i, z_j)], \tag{3.1}$$

where α is a normalizing factor to ensure the integral to unity, $\mu(z_i)$ are functions modeling the first order property, and $\rho(z_i, z_j)$ are functions representing the pairwise interaction. Usually, for stationary point process, $\mu(z)$ is set to be a constant β for all points, while defining $\rho(z_i, z_j)$ as follows:

$$\rho(z_i, z_j) = \begin{cases} 1, & \| z_i - z_j \| > r \\ \gamma, & \| z_i - z_j \| \leq r \end{cases}. \tag{3.2}$$

Then the PDF is simplified to be:

$$f(\mathbf{z}) = \alpha \beta^{n(\mathbf{z})} \gamma^{p(\mathbf{z})}, \tag{3.3}$$

where $p(\mathbf{z})$ is the number of point pairs that are less than r units apart in distance, and α, β, $0 \leq \gamma \leq 1$ are all constants. If $\gamma = 1$, there is no interaction between points, and it can be simplified to a PPP with intensity β. So the Gibbs processes include PPP as a special case. With different assignments for the parameters β and γ, there are different kinds of pairwise interaction processes, such as the Strauss process, Hardcore process and Geyer process. We will give brief description on these point processes as follows.

The Poisson Hardcore Process

A hardcore point process is a kind of point process in which the constituent points are forbidden to lie closer than a certain positive minimum distance. Compared to other hard-core processes, PHCP has the promising merit of fitting efficiency. By setting $\gamma = 0$ in Eq. (3.2), the PDF of Poisson hard-core process can be written as:

$$f(\mathbf{z}) = \alpha\beta^{n(\mathbf{z})}\mathbf{1}(p(\mathbf{z}) = 0), \tag{3.4}$$

The indicator function in the above equation is 1 if the pair number $p(\mathbf{z})$ is 0. Intuitively, the probability density is zero when any pair of points is closer than r units.

The Strauss Process

Strauss point process constitutes a large part of Gibbs processes, and specifically it is a model for characterizing spatial inhibition if the parameter γ ranges from 0 to 1. Its PDF is similar to Eq. (3.3), where each point contributes a factor β to the probability function, and each pair of points closer than r units contributes a factor γ. For the two marginal values of γ, $\gamma = 1$ reduces the Strauss process to a PPP, while $\gamma = 0$ makes it to be a hard-core process as mentioned above.

The Geyer Saturation Process

The Geyer process is a generalization of Strauss process, which is also able to model the clustering effect of a point pattern by tuning the parameter γ. Actually, as seen in Eq. (3.3), the probability density is not integrable if $\gamma > 1$, which is essential for modeling clustering effect. In order to make the PDF integrable, a saturation threshold is added and the PDF becomes:

$$f(\mathbf{z}) = \alpha\beta^{n(\mathbf{z})}\gamma^{\min(p(\mathbf{z}),sat)}. \tag{3.5}$$

Due to the presence of sat, the increasing trend of the PDF when $\gamma > 1$ is limited thus makes the model capable to characterize clustering effect. Moreover, the Geyer saturation process will reduce to a PPP for $sat = 0$, or a Strauss process for $sat \to \infty$.

3.3.3 Clustered Point Processes

Among clustered point processes where attraction occurrs between nearby nodes, Neyman-Scott processes are special examples of Poisson cluster processes [20], which are commonly used in spatial statistics. The points following Neyman-Scott processes consist of the set of clusters of offspring points, centered around an unobserved set of parent points. The parent points form a homogeneous Poisson process of intensity λ_p, while the offspring points around per cluster are random in the number and are scattered independently with identical spatial probability density around the origin. The Matern cluster process (MCP) and Thomas cluster process (TCP) are two representatives of Neyman-Scott processes, and they are distinguished by the difference on how the offspring points are distributed around the cluster center.

Matern Cluster Process

Matern cluster process is a special case of the Neyman-Scott process, where the number of offspring points per cluster is Poisson distributed with intensity λ_c, and their positions are placed uniformly inside a disc of radius R centred on the parent points. We assume that the cluster centers form the point pattern $\mathbf{c}$ which is Poisson distributed with intensity $\lambda_p > 0$. For $\mathbf{c} = \{c_1, c_2, ..., c_n\}$, associate each c_i with a PPP $\mathbf{z}_i$ with intensity $\lambda_c > 0$ and these offspring point processes are independent with each other. The density function at a point ξ around parent point c_i can be written as:

$$f(\xi - c_i) = \frac{2r}{R^2}, \qquad \text{for} \quad r = \parallel \xi - c_i \parallel \leq R. \tag{3.6}$$

Thomas Cluster Process

Unlike the uniform spatial distribution of offspring points around the parent points in MCP, the isotropic Gaussian displacement is utilized in TCP. Replacing the corresponding parameter R in MCP, a standard deviation of random displacement of a point from its cluster center marked as σ is adopted along with the densities λ_p and λ_c. Then the density function of TCP is:

$$\begin{aligned} f(\xi - c_i; \sigma^2) = \frac{1}{2\pi\sigma^2} exp[-\frac{1}{2\sigma^2} \parallel \xi - c_i \parallel^2], \\ \xi \sim N(c_i, \sigma^2). \end{aligned} \tag{3.7}$$

MCP and TCP are widely used in the spatial modeling of aggregated distribution phenomenon. Considering the convenience [63] and tractability [35], both MCP and TCP are employed as cluster point processes models to characterize BS locations here.

3.4 Fitting Methods and Evaluation Statistics

Given the real data ready to be analyzed and various point processes as candidates for accurate modeling, appropriate statistical analysis is essential to connect these two components. Similar with the common statistical estimator based on observed values, the maximum likelihood estimation method is straightforward and very powerful here. Using likelihood-based method (pseudo-likelihood and composite likelihood), the most appropriate parameters are obtained for each point process by fitting to the observed point pattern. Afterwards, relevant evaluation statistics are calculated for each fitted model and compared to that of the real point pattern, in order to identify which point process is the most suitable model for the real BS locations.

3.4.1 Fitting Methods for Point Processes

Likelihood-based fitting method is a common fitting approach in stochastic geometry. Combined with the probability density description of Gibbs point processes, the method of maximum pseudo-likelihood is direct and very convenient for fitting and obtaining the corresponding parameters.

Maximum Pseudolikelihood Method

For PPP fitting process, the method of maximum pseudo-likelihood is the same as maximum likelihood approach. For a spatial point pattern $\mathbf{z}$ observed in a bounded region W, the homogeneous Poisson point process with intensity $\lambda > 0$ has a likelihood function $f(\mathbf{z};\lambda) = exp\{-(\lambda-1)\}\|W\|\lambda^{n(\mathbf{z})}$, where $n(\mathbf{z})$ denotes the number of points in $\mathbf{z}$ and $\|W\|$ is the volume of W. This yields the maximum likelihood estimate $\tilde{\lambda} = n(\mathbf{z})/\|W\|$.

For Poisson hard-core process, r can also be obtained by the method of maximum pseudo-likelihood. In the fitting process, different values of r are tested and then we obtain the corresponding fitted models by the maximum pseudo-likelihood method and select the value of r whose fitted model has the largest maximum pseudo-likelihood. Similarly, the other parameters in Eq. (3.4) can be obtained by using this method again.

For Strauss point process, whose density function is defined in Eq. (3.3), there are four parameters to be determined, namely regular parameters α, β and γ along with the irregular parameter r which is the interaction radius. Firstly, r is selected from the empirical range $[R/2, 4R]$ by the method of maximum profile pseudo-likelihood, where R is the average distance to the nearest neighbor of each point in the point pattern $\mathbf{z}$. Then, after the irregular parameter r is obtained, the other regular parameters can be determined by the maximum pseudolikelihood method repeatedly.

The fitting procedure for Geyer point process is similar to that of the Strauss process, except that another irregular parameter *sat* is added. Usually, the range of *sat* is chosen to be relatively lower in order to make the evaluation of the pseudo-likelihood computationally fast, like [1, 5] in this chapter. All the fitting and simulation processings are completed with the *Spatstat* package in *R* language [7].

Composite Likelihood Approach

The pseudo-likelihood method is too computationally intensive to be applicable for Neyman-Scott point processes. Composite likelihood approaches have been proposed as efficient and feasible ways to deal with this problem, and they can be performed for any process with a second-order intensity function [20]. The second-order statistics of Neyman-Scott point process are well defined. Thus, the statistical properties of MCP and TCP match very well with the fitting process of composite likelihood approach. Concretely, the composite likelihood is firstly formed by introducing some pairwise composite likelihood functions that are defined by second order statistics of the underlying process, and then used for estimating the unknown parameters. The estimation process is computationally simple and can provide consistent results [39]. So in this chapter, in order to be consistent with the pseudo-likelihood method in Gibbs point processes modeling, we adopt composite likelihood method to fit the Neyman-Scott point processes to the real data sets.

3.4.2 Goodness-of-Fit Evaluation Statistics

After the fitting procedure, the goodness of fitting results is verified using some evaluation statistics. There are many statistics being able to characterize the distribution of a point pattern, such as the pairwise correlation function $g(r)$ and the Besag-Ripley's L-function [20]. Indeed, as we are analyzing the spatial structure of BS locations in cellular networks, the practical network performance metric can also be introduced as a reasonable reference for evaluation. In this chapter, the classical statistics like L-function and network performance metrics like coverage probability are employed as evaluation statistics in the identification of different point process models.

Mathematical Metrics from Point Process Theory

In stochastic geometry theory, second-order statistics on spatial point processes describe the so called average behaviour of the point process of interest and give information on many scales of distance. Ripley's K-function is one of the widely used second-order statistics to characterize a point process. Concretely, it is related to point location correlations and can be defined as:

$$K(r) = \frac{1}{\lambda}\mathbb{E}[\Phi(\mathbf{z} \cap B(x,r)\backslash\{x\})|x \in \mathbf{z}], \tag{3.8}$$

where λ is the intensity, $\Phi(\mathbf{z})$ is the number of points in $\mathbf{z}$ and $B(x,r)\backslash\{x\}$ represents the circle centered at x with radius r while eliminating point x. $\lambda K(r)$ can be interpreted as the mean number of points $y \in \mathbf{z}$ that satisfy $0 < \|y - x\| \leq r$, given $x \in \mathbf{z}$.

L-function is a transformation of the Ripley's K-function, which is widely used to test the validity of a point process [75]. It reflects the regularity or clustering property of a point pattern and is defined as:

$$L(r) = \sqrt{\frac{K(r)}{\pi}}. \tag{3.9}$$

For a completely random (uniform Poisson) point pattern, the theoretical value is $L(r) = r$, which is used as a baseline to judge a point pattern's spatial characteristic [75]. If $L(r) < r$, then there is dispersion on this r scale and should be modeled by a repulsive point process; otherwise it is aggregated if $L(r) > r$ and should be modeled by a clustering point process. Due to its explicitness and importance, L-function is adopted as the basic statistical metric in this chapter.

Service Performance Metrics in Cellular Network

In order to find a realistic model, we choose the coverage probability as an evaluation metric to bridge the modeling validity and actual network performance. More formally, the coverage probability of a specific region is the probability that the signal to interference ratio (SIR) of a randomly located user achieves a given threshold in the surrounding cellular networks. Assuming each mobile user connects to the BS that offers the highest received power, while the other BSs in the region transmit as interferers as the frequency reuse factor is assumed to be 1. Apparently, the SIR of each user and the resulting overall coverage probability depend on the transmit powers of the BSs, the random radio channel and the path loss propagation. Randomly selected in the region of $\mathbf{z}$, the resulting received SIR in position s is calculated as:

$$\mathbf{SIR}(s, \mathbf{z}) = \frac{P_y h_y d(s,y)^{-\alpha} s_y}{\sum_{x \in \mathbf{z}\backslash y} P_x h_x d(s,x)^{-\alpha} s_x}. \tag{3.10}$$

P_x and P_y are the transmitted powers of the corresponding interfering BSs and serving BSs and Rayleigh fading is adopted as h_x, $h_y \sim \exp(1)$[‡]. s_x, s_y reflect the shadowing effect and is modeled as log-normal distribution. The path loss exponent α is assumed to be 4 for dense urban scenario and 2.5 for rural regions.

[‡]Exponential distribution with parameter 1.

To identify whether a point process model is suitable for a point pattern or not, we firstly fit these introduced models to the specific sample, then get proper parameters for each model using likelihood-based method mentioned in previous subsection. After that, the critical envelopes are set up as follows. Firstly, we calculate the theoretical mean value of the summary statistic of a fitted model. Then, 199 realizations of each fitted model are generated. For each simulation, we compare the simulated curve to the theoretical curve and compute the maximum absolute difference between them (over the r distance scale or SIR threshold). This gives a deviation series value for each of the 199 simulations. Finally, we take the 10th largest of the deviation value and call it dev. Then the simultaneous envelopes are of the form $low = expected - dev$ and $high = expected + dev$ where $expected$ is either the theoretical value (PPP) or the estimated theoretical value (other models). These simultaneous critical envelopes have constant width $2 * dev$ and reject the null hypothesis if the curve of the desired evaluation metric lies outside the envelope at any value of the r or SIR. This test has an exact significance level $\alpha = 10/(1 + 199) = 5\%$ [7].

3.5 Fitting Results: Case Studies and Large-scale Identification

In this section, as a case study, we first perform the fitting and hypothesis testing for the two small regions in Figs. 3.3 and 3.4 to describe the identification procedure clearly. Specifically, for the dense urban area, separate spatial characterization is applied to both macrocells and microcells and the accuracies of respective models are testified. After the sample analysis, we conduct the large-scale identification across the whole province areas and obtain the outage probability of each candidate point process that models the randomly chosen regions in terms of L function (see Eq. 3.9).

3.5.1 Case Studies for Different Scenarios

In this subsection, we will conduct the case studies for both urban and rural areas. Besides, in urban scenario, we separate the macrocells and microcells into different point patterns, and put forward the modeling process for each of them. Based on these case studies, we can have a brief review on the modeling process of spatial distribution, including the model selection, fitting procedure and performance evaluation. On the other hand, the fitting results of case studies shed light on the BS deployment difference between urban and rural areas, between macrocells and microcells.

Spatial Pattern of BSs in Urban Areas

For the dense urban region in Fig. 3.3, all BS locations constitute point pattern $\mathbf{x}$. Respectively, the 84 macrocells are referred as point pattern $\mathbf{x_1}$ and the microcells make up point pattern $\mathbf{x_2}$. Before the point processes fitting, the L function of the three point patterns are measured and depicted in Fig. 3.6.

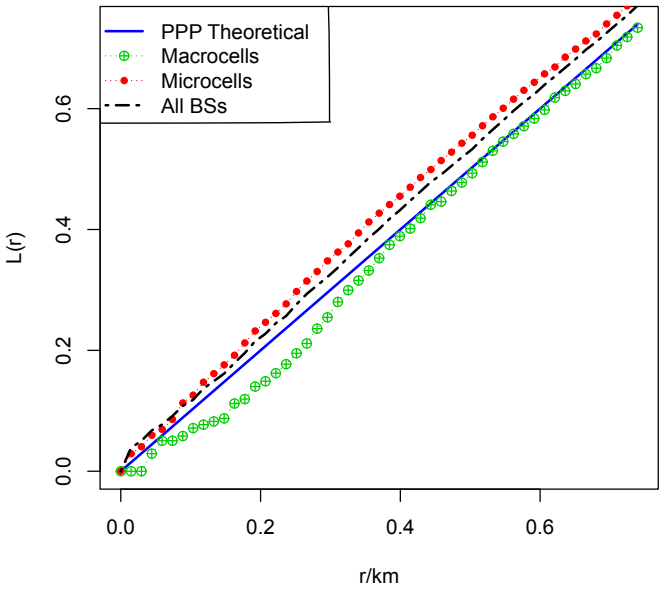


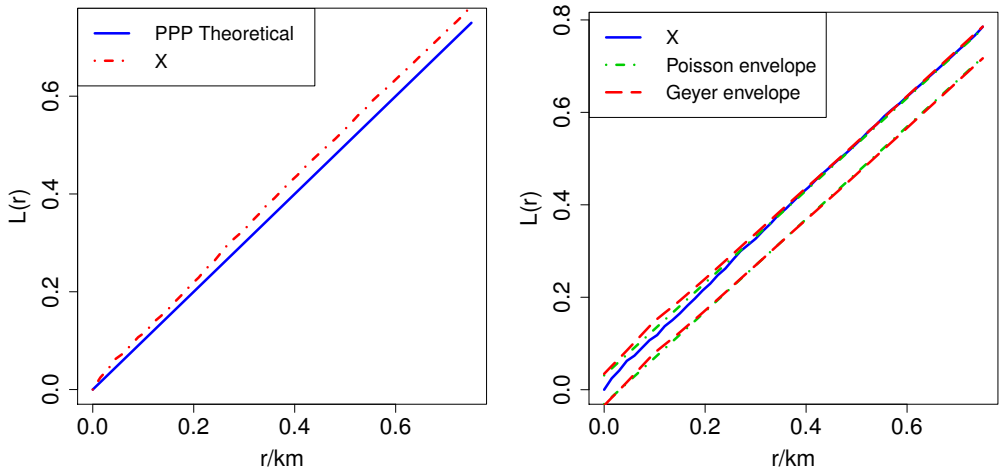
FIGURE 3.6: L function of point pattern $\mathbf{x}$ as all BSs, subset $\mathbf{x_1}$ as macrocells and subset $\mathbf{x_2}$ as microcells, compared with the theoretical curve for PPP.

From Fig. 3.6, we find that the L function curves of both point patterns $\mathbf{x}$ and $\mathbf{x_2}$ are above the theoretical curve of PPP. It means that the whole set of BSs ($\mathbf{x}$) in this region appears to be clusteringly distributed, and so does the microcells' subset $\mathbf{x_2}$. On the other hand, the L function shows that the macrocells' subset $\mathbf{x_1}$ is repulsively deployed because the curve is clearly below the theoretical curve.

Next, we will conduct the modeling processes separately for macrocells and microcells, i.e. point pattern $\mathbf{x_1}$ and $\mathbf{x_2}$. Since they are just subsets of the overall BSs in this region, the network performance metric is not considered. Thus for simplicity, the spatial structure of these detached BSs is only verified here by applying the L function statistics. For the whole BSs set, both L function and coverage probability are utilized as evaluation metrics to test the fitness of various candidate models.

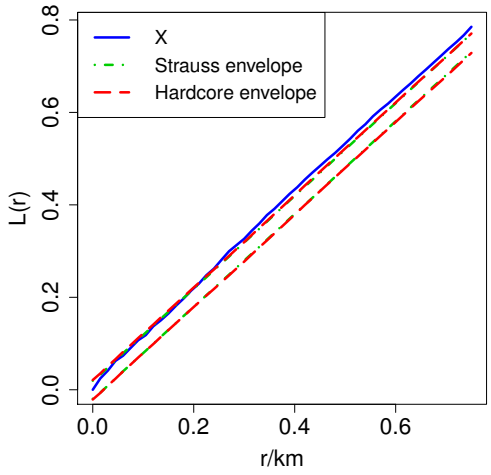
Spatial Modeling for All BSs

Before separate modeling for macrocells and microcells, the spatial distribution of $\mathbf{x}$ is investigated here. The spatial structure of BSs in dense urban area gives an indirect vision of mobile users and traffic demand in cellular networks. In this part, we use both metrics (L function and coverage probability) to test which model is suitable for the spatial pattern of $\mathbf{x}$.

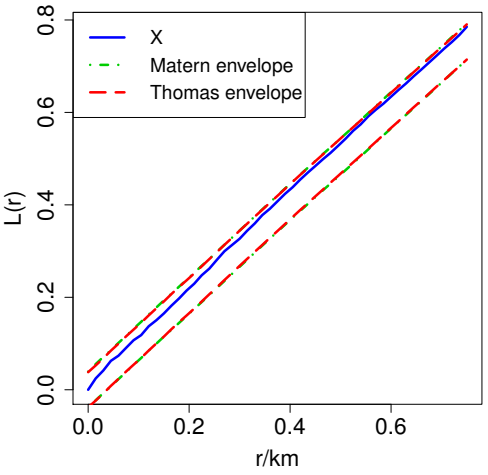


(a) L function of point pattern $\mathbf{x}$.

(b) Poisson and Geyer envelopes.



(c) Hardcore and Strauss envelopes.

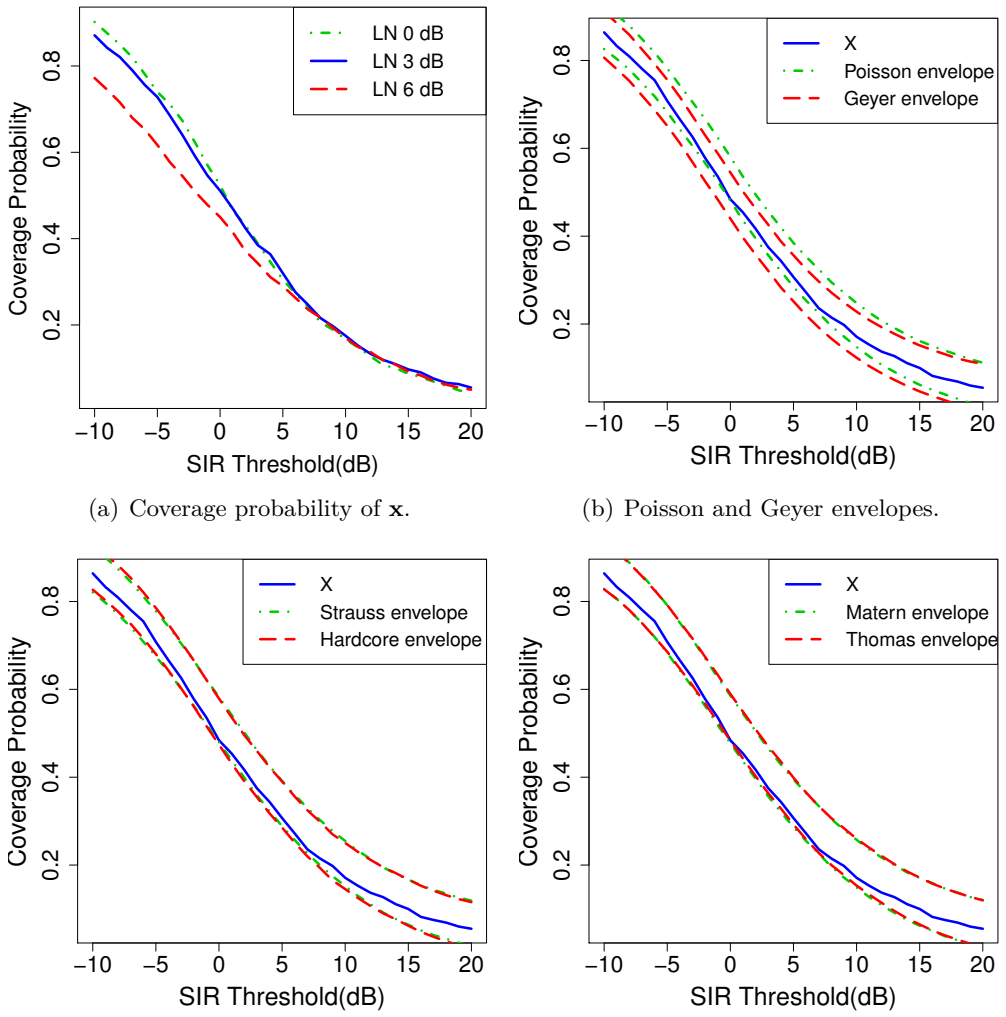


(d) Matern and Thomas envelopes.

FIGURE 3.7: L function of $\mathbf{x}$ and its envelopes of the fitted models.

In Fig. 3.7, the L function curve of point pattern $\mathbf{x}$ and its fitted envelopes are presented. As seen in Fig. 3.7(a), the L function curve of $\mathbf{x}$ is firmly above the theoretical curve of PPP $L(r) = r$, which means that the BSs are aggregately deployed in this region. For the fitted models in Fig. 3.7(b), the curve overflows the envelope of the fitted PPP and Geyer process thus rejects these two model hypotheses. The same result is shown for Strauss and Hardcore in Fig. 3.7(c), and for MCP and TCP in Fig. 3.7(d). All of the high bounds of the fitted envelopes can not

surround the real curve, which means that this sample region is too aggregately distributed to be captured by these six point process models.



(a) Coverage probability of $\mathbf{x}$.

(b) Poisson and Geyer envelopes.

(c) Hardcore and Strauss envelopes.

(d) Matern and Thomas envelopes.

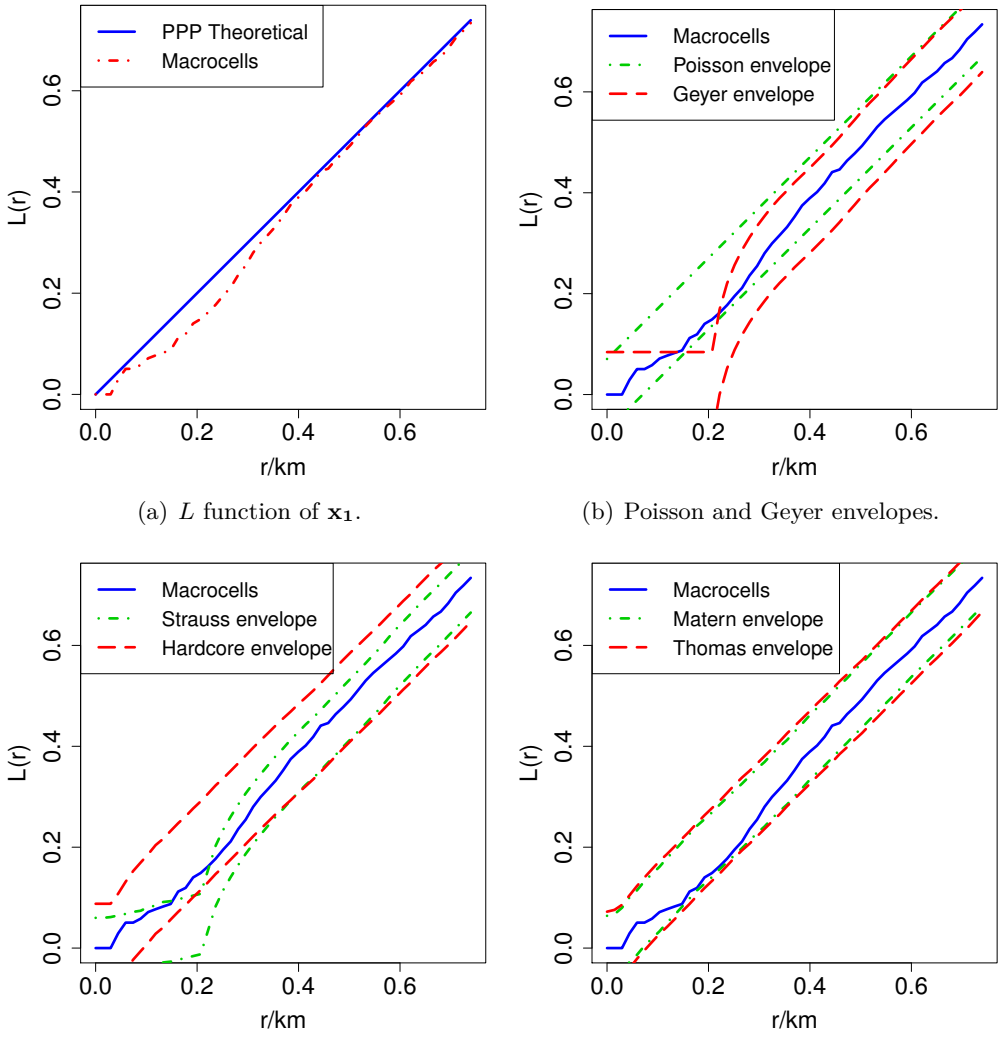
FIGURE 3.8: Coverage probability of $\mathbf{x}$ and its envelopes of the fitted models.

Besides the L function, the identification results of another metric (i.e. coverage probability) are presented in Fig. 3.8. Firstly, the coverage probability of point pattern $\mathbf{x}$ with different lognormal shadowing parameter is depicted Fig. 3.8(a). Then for the fitted models, lognormal shadowing of 3dB is adopted to calculate each envelope. We can observe that coverage probability is not distinguishable in the modeling hypotheses testing since the envelopes of each candidate model surround that of the real data very well.

Spatial Modeling for Macro BSs

For the subset point pattern $\mathbf{x_1}$, since macro BSs are deployed to satisfy coverage requirement, the points tend to be neither too close nor too far away from each other, as seen in Fig. 3.3. To

describe this property explicitly, we fit the six candidate models introduced above to the point pattern $\mathbf{x_1}$, and plot the envelopes of L function of these fitted models.



(a) L function of $\mathbf{x_1}$.

(b) Poisson and Geyer envelopes.

(c) Hardcore and Strauss envelopes.

(d) Matern and Thomas envelopes.

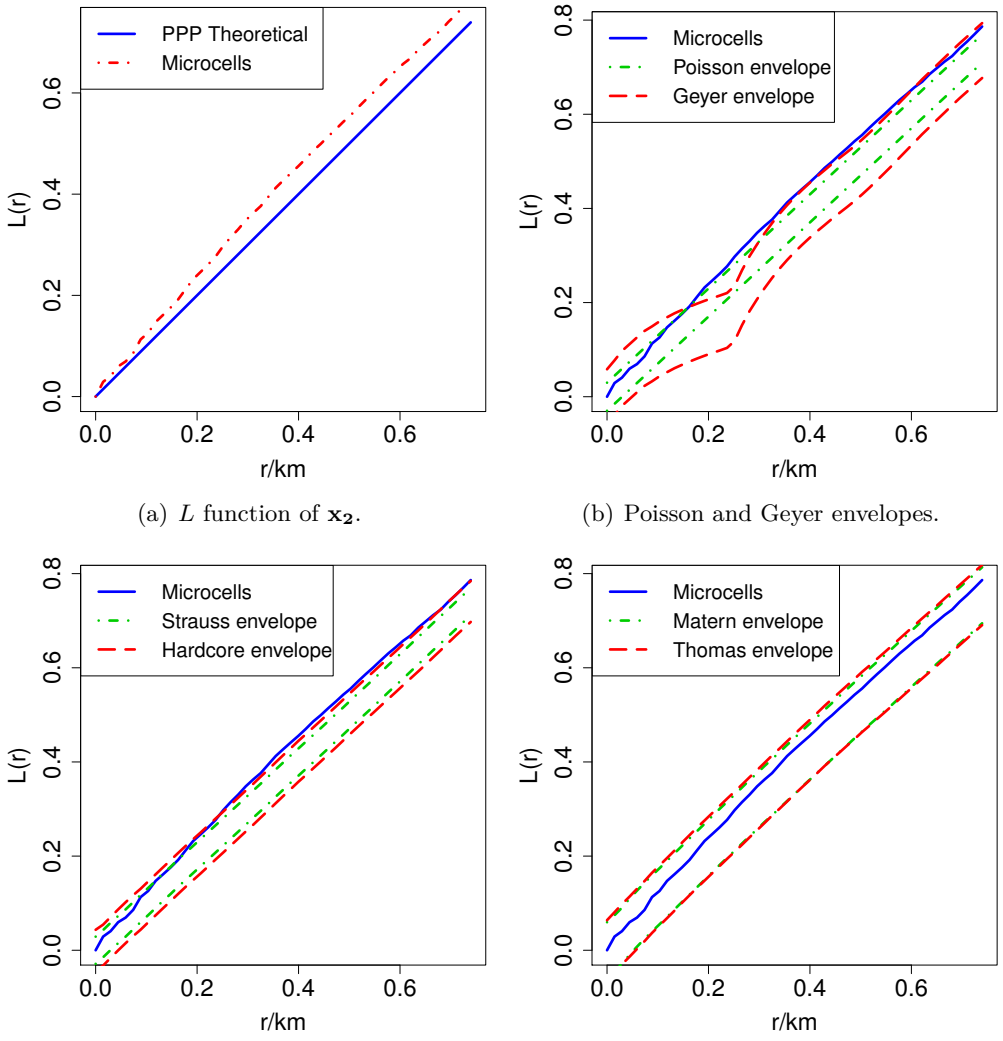
FIGURE 3.9: L function of $\mathbf{x_1}$ and its envelopes of the fitted models.

The L function of $\mathbf{x_1}$ (macrocells) is depicted in Fig. 3.9 along with the envelopes of its fitted point process models. As seen in the Fig. 3.9(a), the L function is exactly below the theoretical curve of PPP, which indicates that the macro BSs tend to be dispersively distributed. Besides it, the envelopes in Fig. 3.9(b) show that the PPP hypothesis for point pattern $\mathbf{x_1}$ cannot be rejected by this metric, while Geyer process is the opposite. It is the same situation in Fig. 3.9(c), we can deny the Strauss hypothesis of $\mathbf{x_1}$ but reserve the Hardcore claim. Surprisingly, the envelopes of the fitted MCP and TCP models capture the real data very well as PPP does.

Remark 3.1. Macro BSs tend to have a repulsive distribution in dense urban area, which reflects its original functionality in cellular networks deployment.

Spatial Modeling for Micro BSs

Unlike macro BSs, microcells are usually deployed by operators to diminish coverage hole and offload heavy traffic from macrocells. As seen in Fig. 3.3, micro BSs are more intensively distributed than macro BSs. Visibly, the L function of $\mathbf{x_2}$ and its fitted envelopes are presented in Fig. 3.10.



(a) L function of $\mathbf{x_2}$.

(b) Poisson and Geyer envelopes.

(c) Hardcore and Strauss envelopes.

(d) Matern and Thomas envelopes.

FIGURE 3.10: L function of $\mathbf{x_2}$ and its envelopes of the fitted models.

Comparatively, the L function of microcells is totally above the theoretical value of PPP which verifies the clustering nature of the distribution of micro BSs. More specifically, in the Fig. 3.10(b), the fitted PPP and Geyer process fail to contain $\mathbf{x_2}$ within their L function envelope. Thus PPP and Geyer process model can be rejected by this hypothesis test, so do Strauss process and Hardcore process in Fig. 3.10(c). These results confirm the aggregation property of microcells' distribution in this selected region. While in Fig. 3.10(d), the L function envelopes of MCP and TCP accept that of $\mathbf{x_2}$ very well. Combining these results above, we can conclude

that the microcells in this dense urban region tend to be aggregately distributed and may be well characterized by MCP and TCP.

Remark 3.2. Micro BSs in dense urban area tend to be aggregately deployed to fulfill the heavy concentrated capacity demand.

Spatial Pattern of BSs in Rural Areas

As seen in Table 3.1, the BSs density in rural regions are much less than urban regions, due to the relatively smaller population and much less service demand. In this subsection, we will turn to the representative sample of rural region to check the difference between the urban and rural BSs deployment, which in return reflects the urbanization process and extent of different regions.

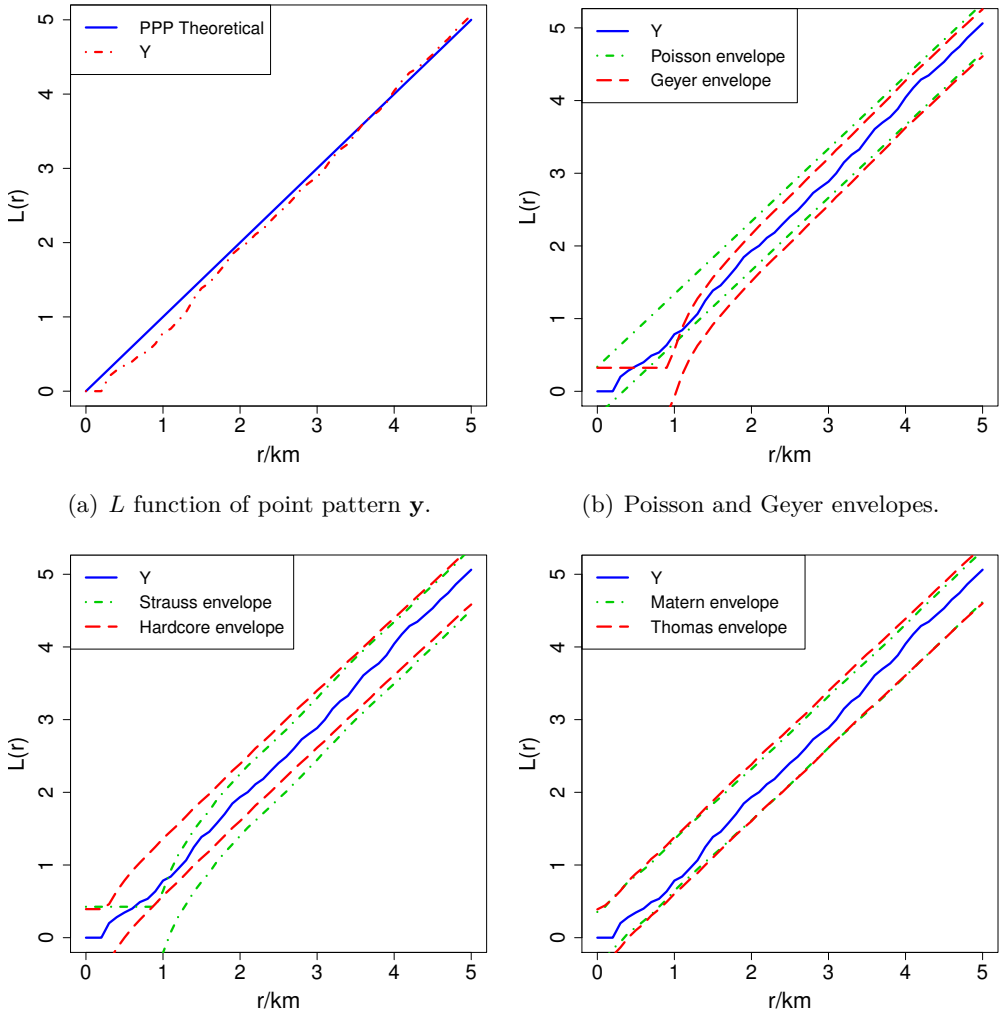
In the selected rural region as illustrated in Fig. 3.4, there are 79 BSs with only 5 microcells within this $20 \times 20\ km^2$ area which is referred as point pattern $\mathbf{y}$. Since the number of microcells is very few, we analyze the whole set of BSs in this region regardless of the different BS types.

In Fig. 3.11, the L function of point pattern $\mathbf{y}$ is presented with envelopes of its fitted point processes. In Fig. 3.11(a), the regularity of point pattern $\mathbf{y}$ is clearly observed, as the L function curve of $\mathbf{y}$ does not exceed the theoretical curve of PPP for the most part. For the fitted models, as in Fig. 3.11(b), the envelope of PPP encompass the L function curve very well while Geyer point process fails in the range near 1 km. Moreover, in Fig. 3.11(c), PHCP captures the curve completely while Strauss process is unsatisfied. However, in Fig. 3.11(d), both of the envelopes of MCP and TCP fit the curve remarkably. This result indicates that the so-called cluster processes can also manage to be applied to the regular point pattern since the parameters of these models have a relatively high degree of freedom.

Besides the L function, the coverage probability of point pattern $\mathbf{y}$ and the corresponding envelopes are also depicted in Fig. 3.12. Counterintuitively as in Fig. 3.8, the envelopes of all fitted models encompass the real curve of $\mathbf{y}$ very well, thus we show that the coverage probability metric is not distinguishable in this test. In this respect, in the following part of large-scale spatial distribution identification, we adopt the L function as the only goodness-of-fit metric to determine the applicability of fitted point processes in regard to huge amount of selected regions.

3.5.2 Large-scale Spatial Modeling Identification

After the modeling procedure of representative regions, we will carry out a large-scale identification, in order to achieve a more comprehensive result for BS locations. Basically, the



(a) L function of point pattern $\mathbf{y}$.

(b) Poisson and Geyer envelopes.

(c) Hardcore and Strauss envelopes.

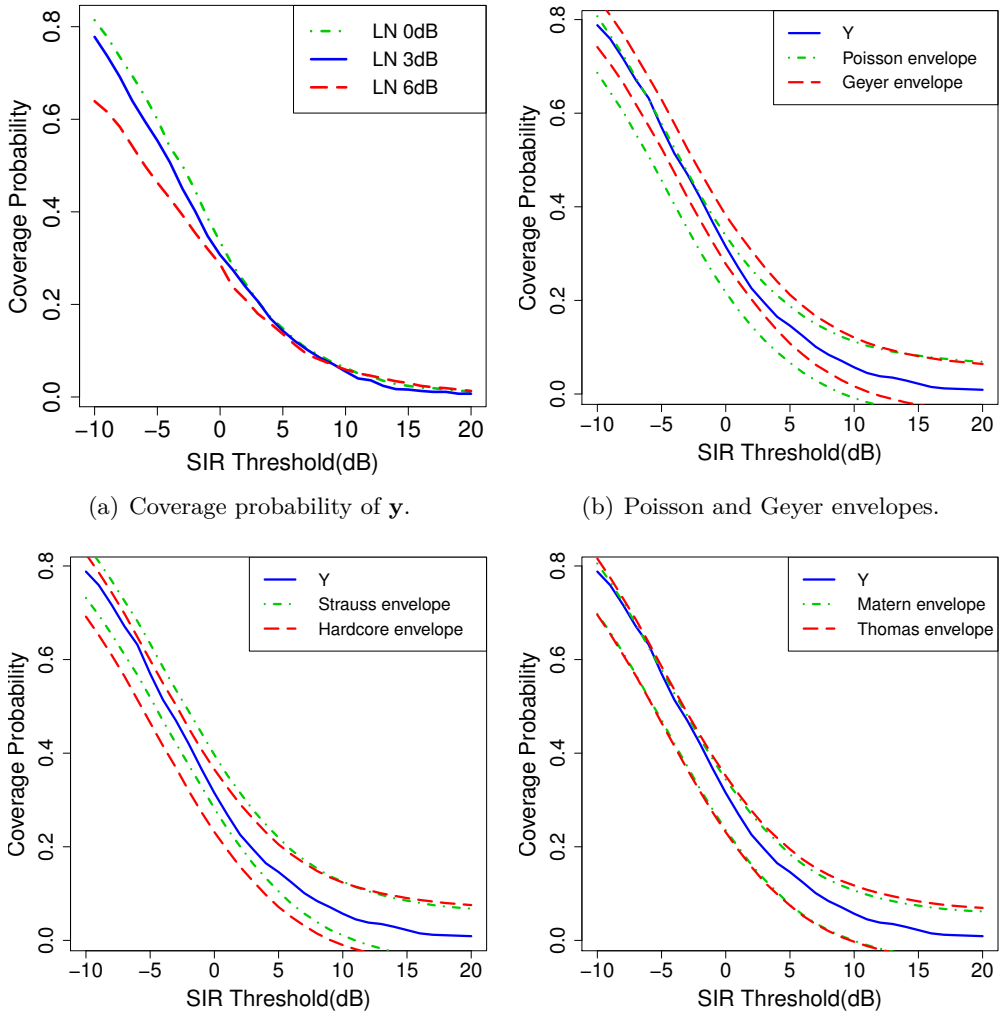
(d) Matern and Thomas envelopes.

FIGURE 3.11: L function of $\mathbf{y}$ and its envelopes of the fitted models.

identification process contains two steps. Firstly, we test the disperse or clustering property of BS locations for all kinds of areas such as rural and urban areas, and for different types of BSs such as macrocells and microcells. Then, after obtaining the spatial characteristics of BSs, we go further to identify the suitable spatial point process for the corresponding scenarios. Similarly, both of these two steps are based on the real data from the same cellular network operator and the L function aforementioned above.

Spatial Characteristics of BSs Distribution

In order to reveal the fundamental spatial characteristics of BSs distribution, the testification of disperse or aggregate property is the first-step procedure, meanwhile it is a straightforward way to verify the accuracy of PPP model as well.



(a) Coverage probability of $\mathbf{y}$.

(b) Poisson and Geyer envelopes.

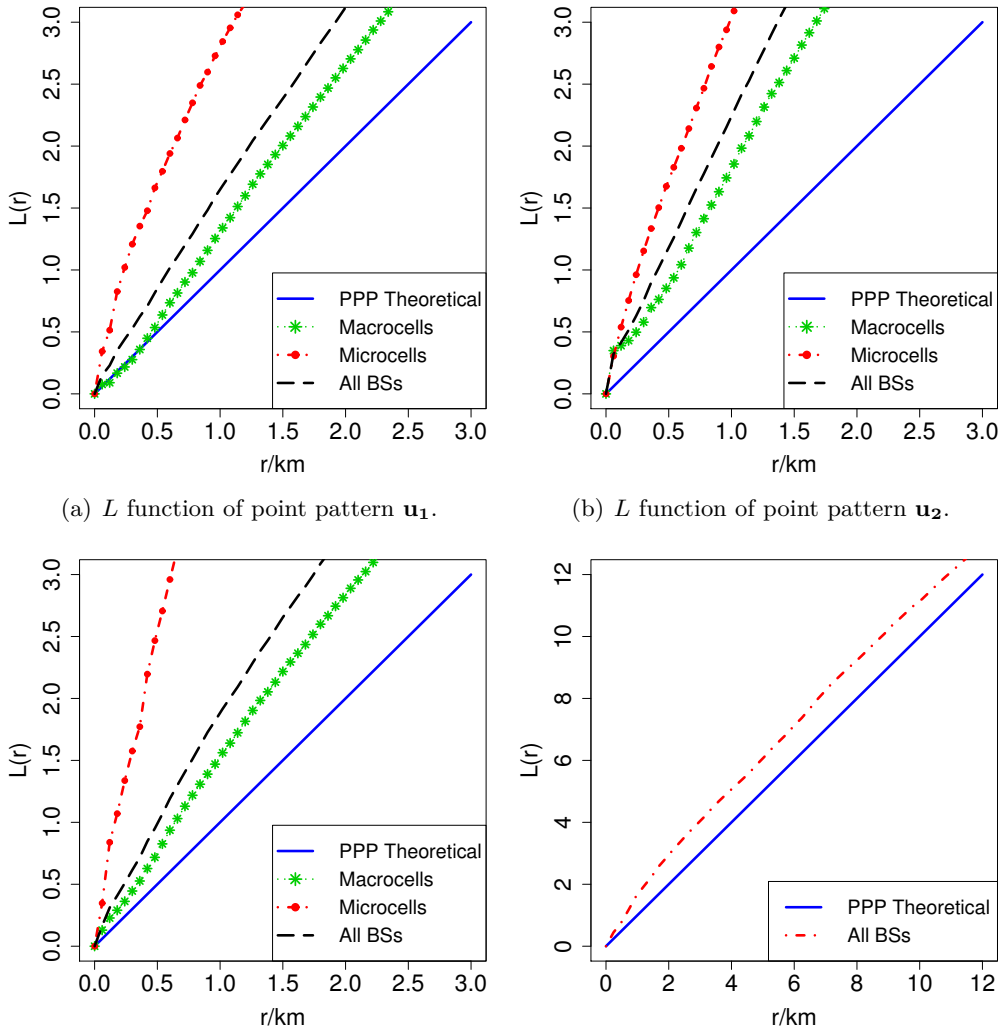
(c) Hardcore and Strauss envelopes.

(d) Matern and Thomas envelopes.

FIGURE 3.12: Coverage probability of $\mathbf{y}$ and its envelopes of the fitted models.

Actually, the L function is computed on a distance scale and it varies depending on the locations of points in the selected region. Specifically, if $L(r) > r$, we say this point pattern is aggregated on this r scale, otherwise we call it dispersed in this distance. Thus, this property (dispersion or aggregation) can be evaluated on the distance scale, rather than on a particular point pattern. According to this methodology, we firstly examine four sufficiently large areas chosen from the real data set, and find the clustering tendency and property of BS locations on the large scale. Moreover as a comparison, we also select thousands of small regions covering urban and rural areas to verify this claim on a smaller scale.

Firstly, the L function of these four large areas are depicted in Fig. 3.13(a-d) respectively. The first three point patterns (i.e. $\mathbf{u_1}$, $\mathbf{u_2}$, $\mathbf{u_3}$) are from urban area of 20×20 km^2 and the point pattern $\mathbf{r_1}$ from rural area is 50×50 km^2. We can observe that, the BSs are aggregately distributed on respective distance scale except that a small number of the macrocells in the area of city A are dispersed in the range of (0,0.3) km distance. Mostly, the L functions of



(a) L function of point pattern $\mathbf{u_1}$.

(b) L function of point pattern $\mathbf{u_2}$.

(c) L function of point pattern $\mathbf{u_3}$.

(d) L function of point pattern $\mathbf{r_1}$.

FIGURE 3.13: The dispersion or aggregation examination of large-scale areas in urban and rural regions.

these areas are far above the PPP theoretical curve, which in turn verifies the inaccuracy of the widely-accepted PPP assumption. So we can conclude that the BSs of cellular networks are generally aggregately distributed in various areas.

Furthermore, after the large-scale testification of the clustering property of BS locations, hereinafter we conduct small scale identification procedure with fine spatial resolution in a probabilistic manner to strengthen this claim. We randomly select 3000 small regions of 6×6 km^2 from the whole coverage areas of the three cities (A, B, C) and 5000 small regions of 20×20 km^2 from the whole rural area. For both kinds of small regions, the investigated distance is assumed to be 0 to quarter of the length of region side, namely (0,1.5) km for the urban regions and (0,5) km for the rural regions. For each distance scale, we compute the corresponding clustering probability (i.e. $P(L(r) > r)$) in the region set, as plotted in Fig. 3.14-3.15.

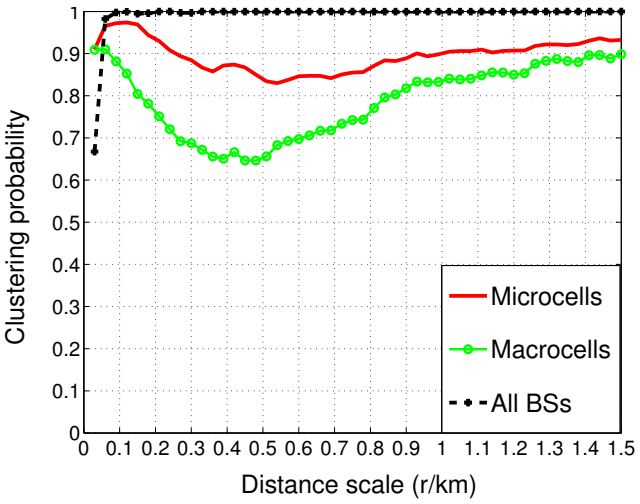


FIGURE 3.14: Clustering probability of BSs on different distance scales in urban regions.

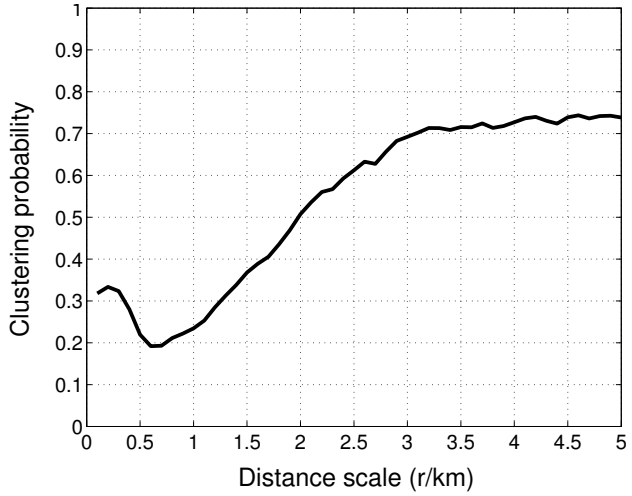


FIGURE 3.15: Clustering probability of BSs on different distance scales in rural regions.

For urban regions, the three probability curves for microcells, macrocells and all BSs are mostly above 0.65, indicating that clustering property is significant on small distance scales as well. Specifically, microcells are more likely to be aggregated than macrocells, but less than their combination (all BSs) whose clustering probability curve is mostly above 0.95. The high probability of clustering effect on small scales in urban regions verifies the conclusion that the BSs tend to be aggregately distributed in urban areas.

For rural regions, as observed in Fig. 3.15, there are more regions which are dispersed than that are aggregated within the distance range of (0,2) km. However, within the range of (2,5) km, the probability of clustering increases with the distance scale. The disparity between different distance scales reflects the evolving complexity of BS locations in rural regions.

Remark 3.3. Conclusively, BSs tend to be aggregately distributed in cellular networks in general. Specifically, the effect of clustering is more significant in urban areas than that in rural areas due to the comparatively higher traffic demand and more densely distributed population.

Point Processes' Accuracy to Model BS Locations

After the description of spatial characteristics of BS locations, we go further to find the suitable point process models for different kinds of BSs and geographical regions in a probabilistic manner. Again, we employ the randomly selected 3000 urban regions and 5000 rural regions as our dataset. For each region in the dataset, we fit the six model candidates as aforementioned to the real data. Then, for each fitted model, we repeatedly conduct the same process as in Section 3.5.1 and estimate the accuracy of the targeted model by the L function statistic. Consistently, the parameters of the test are the same as in Section 3.5.1, so we build up a hypothesis test with significant level 5%. If the $L(r)$ function curve of real data is out of the envelope bound on any r distance scale, we claim the inaccuracy of this model for modeling BS locations in this specific region. In the test set, we introduce the outage probability of a point process model which is the ratio of the summed number of the respective regions with non-accurate modeling to the total number of the tested regions. As follows, for all pairs of area and model, we present the outage probability in Table 3.2.

TABLE 3.2: Outage probability of different models for modeling BS locations.

Region	PPP	PHCP	Strauss	Geyer	MCP	TCP
City A	79.2%	98.8%	97.0%	91.2%	37.1%	33.7%
City B	81.8%	100%	98.3%	92.8%	53.5%	55.4%
City C	82.8%	99.1%	97.8%	94.6%	41.6%	31.5%
Rural	55.1%	98.3%	99.5%	93.4%	42.6%	30.9%

From Table 3.2, we can observe that because of the clustering tendency of BSs deployment, the accuracy of Gibbs processes (PHCP, Strauss and Geyer) is very low. Concretely, for the three urban areas, the outage probability is approximately 100% for PHCP, over 95% for Strauss, and over 90% for Geyer point process. The average outage probability of the three models is increasing with their partiality to repulsive property which coincides with the clustering nature of BS locations in urban areas. Moreover, the outage probability of PPP is close to 80% for urban areas although being relatively better in rural area (55.1%). On the other hand, the accuracy of Neyman-Scott processes are much better, and the average outage probability is around 40% for both MCP and TCP in urban areas. These results further identify the clustering property of BS locations in urban areas, and particularly verify the inaccuracy of PPP's usage for spatial modeling of BSs in cellular networks.

Meanwhile, we calculate the outage probability of different models for macrocells and microcells separately in urban areas as shown in Tables 3.3-3.4.

After the separation of macrocells and microcells, PPP model has slightly better performance to model macrocells since the clustering effect is less significant. The outage probability of Gibbs processes generally decreases comparing to the mixed BSs case but it is still too high to

TABLE 3.3: Outage probability of different models for modeling macro BS locations.

Region	PPP	PHCP	Strauss	Geyer	MCP	TCP
City A	69.4%	98.8%	93.8%	85.0%	63.0%	61.1%
City B	90.6%	98.9%	96.8%	86.0%	79.4%	79.0%
City C	77.4%	97.7%	96.7%	89.7%	48.0%	39.0%

be adopted. Surprisingly, the accuracy of Neyman-Scott processes gets worse which challenges their suitability of usage in single-tier modeling of macrocells in cellular networks. Nevertheless, it is still reasonable that either MCP or TCP is a better choice for modeling macrocells compared to the other models.

TABLE 3.4: Outage probability of different models for modeling micro BS locations.

Region	PPP	PHCP	Strauss	Geyer	MCP	TCP
City A	99.5%	96.8%	97.0%	89.5%	67.8%	66.1%
City B	99.4%	91.8%	98.9%	94.5%	90.5%	88.1%
City C	97.8%	90.9%	96.4%	83.0%	62.7%	66.5%

For microcells, the outage probability of PPP model is extremely high with average value around 99%, which strongly shakes the common sense of complete randomness in higher tier BSs deployment. Consistently, the outage probability performance of other models is similar with that in macrocells modeling which is inevitably too high. Although the cluster processes MCP and TCP are relatively more accurate than the Gibbs point process models, they are not qualified to model micro BS locations anymore, which clearly implies that some other new models are necessary to characterize the strong clustering property of micro BSs.

In summary, among the commonly used six spatial models including repulsive and clustered point processes, the Neyman-Scott point processes (MCP, TCP) have better accuracies in modeling BS locations. But due to the complexity of actual BS deployment and geographical diversity, neither model is perfectly qualified to reproduce the real scenario in our analysis. Surely, these large-scale identification results give us a broader view on this topic and suggest us to further search more accurate and realistic models for spatial distributions of BSs in cellular networks.

3.6 Conclusion and Discussion

In this chapter, we conducted a large-scale identification on the spatial modeling of BS locations in cellular networks. Based on a large amount of real data from the on-operating BSs, our conclusions are given as follows.

Firstly, we investigated the accuracy of PPP's usage in modeling BSs spatial distributions, and verify that the complete randomness property of PPP model is not valid in on-operating well-developed cellular networks. This result will obviously challenge the rationality of networking performance characterization based on the overwhelming PPP assumption in heterogeneous cellular networks. Secondly, the clustering nature of BSs deployment was discovered and the diversity between macrocells and microcells is exhibited indicating that high tiers (microcells) tend to be more aggregately deployed than lower tiers (macrocells). At last, we showed that the two typical clustering models (MCP and TCP) have improved modeling accuracy but are still not qualified to accurately reproduce the practical BSs distribution scenario, due to the complexity of actual BS deployment and geographical diversity. In this situation, it's necessary to step back from two-dimensional spatial modeling to one-dimensional spatial density of BSs in cellular networks. Next chapter will focus on this topic.

Nevertheless, there is still a dilemma between either adopting a more tractable but less accurate model or employing a practical but intractable model. Meanwhile, more real data from other countries are definitely necessary to identify the universal BS spatial distribution pattern. From these points of view, there are still a lot of future work on this issue to capture the heterogeneous cellular networks evolution.

Chapter 4

Clustering Nature in Cellular Networks: Connecting the Dots

Contents

As reviewed in Chapter 2, clustering phenomenon is widespread in cellular networks, such as the spatial-temporal distribution of the traffic demand and the spatial distribution of BSs. Clustering means that mobile users or network deployment tend to generate traffic or be located in a similar time or space. That's to say, the densities of traffic demand and network deployment are highly skewed temporally or spatially, which draws forth the necessity of characterizing the densities of different metrics. Above that, it's also very important to dig into the cause of these clustering phenomenon and uncover the relationship between different phenomenons. This will be the main topic of this chapter.

In detail, this chapter is divided into six sections. Firstly, we will give a brief introduction on this topic in Section 4.1, including the background and related works, followed by our approach and contributions. After that, the mathematical preliminary will be presented in Section 4.2, where different candidate distributions for characterizing the clustering nature of traffic and BS deployment are introduced. Then in Section 4.3, 4.4 and 4.5, we try to analyze and find the most appropriate models for the spatial density of BS and user traffic demands, and the temporal distribution of mobile instant message (MIM) as examples, respectively. Finally, the discussion and conclusion are given in Section 4.6.

4.1 Introduction

User, traffic and BS are the three fundamentals of RAN in cellular networks. In detail, mobile users generate data requests, and send them to connected BS, and BS replies through the air interface. Therefore, in order to characterize the cellular networks more accurately, it's necessary to investigate the statistical properties of mobile users, traffic demand and BS deployment, both temporally and spatially.

4.1.1 Background

Cellular networks are becoming an inevitable data pipe for diverse mobile devices to access intense contents on the Internet. Understanding how BSs are spatially deployed could prominently facilitate the performance analyses of cellular networks, as well as the design of efficient networking protocols. For example, Poisson distribution is widely adopted to characterize the spatial distribution of BSs and leads to a tractable approach to calculate the coverage probability and traffic rate in cellular networks, by taking advantage of a PPP based theory (i.e., stochastic geometry) [4, 42]. On the other hand, the actual deployment of BSs is highly correlated with human activities in the long term [107, 110]. Humans tend to live together, and their social behaviors would lead to traffic hotspots [107], thus causing BSs to be more tensely

deployed in certain areas as clusters. Furthermore in a wider view, according to the assumption named "preferential attachment" [8], Barabási *et al.* argued that many large networks grow to be heavy-tailed. This claim and reasoning make heavy-tailed distributions appear to be more suitable to characterize the spatial density of clusteringly distributed BSs.

On the other hand, traffic demand variation is another important issue in cellular networks, since it's highly related to the interference management and resource allocation strategies. More importantly, the energy efficiency of communication systems is directly determined by the traffic volume level of the coverage area, which highlights the urgency of fine-grained characterization of mobile user traffic in cellular networks. from these points of view, it's important and necessary to take an in-depth investigation of the traffic demand in the widespread cellular networks, on both temporal and spatial perspectives. Specifically, the spatial density is a good estimator of the two-dimensional distribution of traffic demand, though losing some of the location information. Therefore, we choose to investigate the traffic density (spatially and temporally) based on real measurements.

Specifically, the probes deployed on BSs are able to collect the total traffic volume which get through it to core networks or mobile users. In this aspect, in order to obtain the density description, we assume that the traffic density within the coverage area of one BS is invariant while distinct BSs may differs on this metric. Furthermore, as the number of small cells is increasing rapidly, the coverage area of each cells is going to be smaller and smaller, which makes our invariant traffic density within one cell reasonable. By this method, we are able to obtain the real traffic density across the cellular networks.

Moreover, the increasing deployment of dense small cells causes the cellular networks topology much more complicated than before. Although there were numerous substantial works about the traffic spatial distribution and BSs deployment, the relevant statistical models derived from the former cellular architectures may not be practical to fully reflect the ongoing network evolution. Therefore, by means of analyzing the intrinsic relationship between BSs density and traffic spatial density, we aim to go beyond the varieties of network facility and obtain a deep-level understanding on the fundamental patterns of cellular network evolution.

In addition to the spatial pattern of BSs and traffic demand, the temporal characteristic of user traffic also plays a key role in the overall performance of cellular networks, for example on the average latency metric. As described in Chapter 2, the temporal distribution of traffic demand also exhibits some extent of clustering effect. However, the present study only reveals this phenomenon but lacks an in-depth statistical characterization of it, which could be very useful for tangible analysis and practical simulations.

4.1.2 Related Works

As presented in Chapter 3, there are a lot of works aiming to find the most appropriate model for spatial distribution of BSs in cellular networks [4, 42]. Recently, the modeling accuracy of PPP has been recently questioned [116]. Consequently, in order to reduce the modeling error between Poisson distributed BSs and the practical case [25], some variants of PPP have been exploited to obtain precise analysis results. Due to the disparity of different data set and distinctive evaluation metrics, the academia can hardly reach a consensus on this topic [22, 107]. To overcome the difficulty of lacking large amount of real data, we collected massive BS locations from an on-operating cellular networks to conduct a large-scale identification process on this issue [115]. However, our conclusions indicates that maybe there is not a universal model for the spatial distribution of BSs in cellular networks, since the scenarios vary for different areas and different types of BSs. Actually, the spatial modeling of BS locations is a trade-off between accuracy and tractability. For example, the PPP has the best tractability for performance evaluation in cellular networks since it provides a close-form description of many key metrics like coverage probability and transmission capacity. While, the completely random assumption of PPP models is clearly unrealistic for practical deployment. Therefore, after revealing these facts, it's more reasonable to consider this problem in a different view. In this chapter, we give up to find a more accurate model on two-dimension and try to characterize the clustering feature of BSs distribution[113] in a more direct way (i.e., the spatial density).

Besides the spatial characterization of BSs, traffic demand is another important issue to be considered and Chapter 2 gives a comprehensive review which reveals the corresponding clustering nature in temporal and spatial domain. However, most of the works dealing with traffic distribution lacks a concrete mathematical description of the degree of clustering property. Moreover, burstiness, long-range dependence (LRD) and heavy-tailed properties of broadband wired network traffic have been discovered, and α-Stable model with the above three features was used in [24], [102]. The latest literature [54], first applied α-Stable distribution to model aggregated traffic traces within BSs in the field of cellular networks. Besides traffic spatial distribution itself, its impact on BSs deployment cannot be ignored. Previous works like [109], adopted saturation model to describe the correlations between the two quantities (i.e., BSs density and traffic density) in urban areas, with a small amount of less BSs and traffic records. This result, however, conflicts with the general awareness that BSs distribution and traffic distribution incline to vary consistently. In real network scenarios, the locations of BSs are usually coupled with the requirements of subscribers that often exhibit group users' behavior [107].

Apart from the spatial characterization, the temporal distribution of traffic demand also exhibits heavy-tailed phenomenons. For example, the traffic volume of an investigated BS in peak time is far more than that in midnight. That's to say, the temporal traffic densities are various and can not be assumed as constant any more. Indeed, due to its apparent importance to the protocol

design and performance evaluation of telecommunications networks, there already exists some works towards the traffic modeling in various networks. In fixed broadband networks, researchers showed that aggregate traffic traces demonstrate strong burstiness and could be modeled with α-Stable models [34, 102].

As mentioned in Chapter 2, the traffic volume on a BS or cell level exhibits a periodic pattern while it also lacks statistical description on the temporal density. On the other hand, the investigation over traffic characteristics in wired Internet revealed heavy-tailed distribution phenomena in services like AIM (AOL Instant Messenger) and Windows Live Messenger [51, 101]. Besides, nowadays MIM emerges as killer application for mobile Internet era, thus takes over large part of the overall traffic in cellular networks and can be chosen as a representative case for the temporal modeling of data traffic. Therefore, it is natural to raise a question, namely which one of the aforementioned models is more suitable for MIM traffic? Meanwhile, it remains doubtful whether cellular networks with distinct characteristics from fixed networks [55] need a totally different traffic model?

4.1.3 Approach and Contributions

In order to statistically describe the clustering phenomenon of cellular networks, firstly we need to choose a proper variable to characterize it. In this chapter, we adopt the spatial and temporal densities to fix this issue. Specifically, we examine the spatial density of BSs across a large selected area, and calculate the average traffic density across the cellular networks on a time basis like one day or one week. Based on the empirical traffic densities, we fit various distribution candidates to show how the BS deployment and traffic demand are spatio-temporally clustered in cellular networks.

In detail, for the first part of this chapter, we aim to re-examine the statistical pattern of BSs, and find the most accurate distribution for the corresponding spatial density. By using a large amount of BSs' real data from on-operating cellular networks, we compare the practical distribution of BSs with various representative candidates, including Poisson distribution and some other heavy-tailed distributions. Interestingly, among the exploited distributions, α-Stable distribution could most precisely fit the actual deployment of legacy BSs, which is also consistent with the traffic distribution in broadband and cellular networks [24, 34]. In other words, the spatial distribution of BSs reflects the basic characteristics of traffic demands from users, and could partially exhibit the nature of human activities. We believe that this new finding could contribute to the understanding of the evolution of cellular networks as well as the relevant society development.

For the second part, we try to characterize the spatial density of traffic demand in cellular networks which is also aggregately distributed, as depicted in Chapter 2. Therefore, we conduct

the same fitting process as BS density distribution, and obtain the most appropriate model for traffic density. Unsurprisingly, the traffic density based on another real data set also reflects heavy-tailed properties, and can be well approximated by α-Stable distribution with different parameters. This result leads us to investigate the statistical relationship between BS density and traffic density in cellular networks. Furthermore, we find that there is kind of linear dependence between these two variables, and the slope of this linear formulation can be adopted as an indicator for the advancement of cellular networks.

In the third part, the temporal analysis of traffic demand is presented, based on a typical example of MIM. Like the spatial density of BSs or traffic demand, we sampled the aggregate traffic of randomly chosen BS on temporal scale. Based on real data, we conduct the fitting process with popular distribution candidates. Once again, the α-Stable distribution gives the best fitting performance among all these distribution candidates. Moreover, the accuracy of power-law and lognormal distribution for packet length and inter-arrival time are verified. Furthermore, we try to illustrate the quantitative coincidence for the aggregate traffic, and find that the general central limit theorem may be a good explanation.

4.2 Mathematical Preliminary

In this section, we will introduce the mathematics for analyzing the spatial and temporal distribution of traffic demand in cellular networks. Since the traffic generated or the infrastructure deployed in cellular networks is kind of reflection of human activities, it's essential to know the basics of human dynamics which is highly connected to the heavy-tailed distribution.

4.2.1 Heavy-tailed Distribution

Actually, heavy-tailed phenomenon is a showcase of general imbalance, while its distribution is more skewed than others. For example, the economist Pareto found that 20% of all people receive 80% of all the income. Technically in probability theory, heavy-tailed distributions have probability densities functions whose tails are not exponentially bounded and the mathematical definition is as follows:

Definition 4.1. The distribution of a random variable X is said to be heavy-tailed if:

$$\lim_{x\to\infty} e^{\lambda x} Pr(X > x) = \infty, \qquad \forall \quad \lambda > 0. \tag{4.1}$$

Heavy-tailed distributions could be widely applied to explain a number of natural phenomena, like in human mobility [37] and the Internet topology [33]. Also for cellular networks which

are highly related to human dynamics, heavy-tailed phenomenon is common for mobile user statistics. Actually, there exists many statistical distributions proving to be heavy-tailed, such as the generalized Pareto (GP) distribution, Weibull distribution, and log-normal distribution belong to one-tailed ones with closed-form PDF.

All these common heavy-tailed distributions are able to indicate very large values in extreme cases, which happens for the BS density or traffic demand in cellular networks. That's the reason why heavy-tailed distributions are adopted here to characterize the statistical properties of BS and traffic.

4.2.2 The α-Stable Distribution

A famous heavy-tailed distribution is the α-Stable distribution, which manifests itself in the capability to characterize the distribution of normalized sums of a relatively large number of independent identically distributed random variables [78]. However, the α-Stable distribution, with few exceptions, lacks a closed-form expression of the PDF, and is generally specified by its characteristic function.

Definition 4.2. A random variable X is said to obey the α-Stable distribution if there are parameters $0 < \alpha \leq 2$, $\sigma \geq 0$, $-1 \leq \beta \leq 1$, and $\mu \in \mathcal{R}$ such that its characteristic function is of the following form:

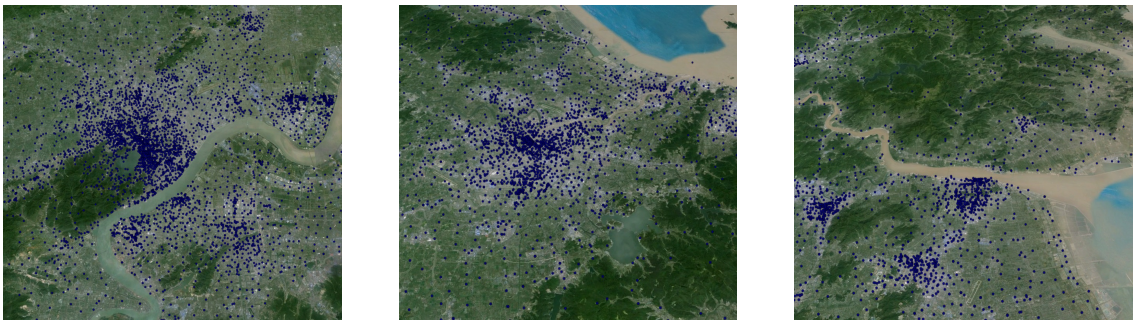
$$\begin{aligned} \phi(\omega) &= E(\exp j\omega X) \\ &= \exp\left\{-\sigma^{\alpha}|\omega|^{\alpha}\left(1 - j\beta(\mathrm{sgn}(\omega))\Phi\right) + j\mu\omega\right\}, \end{aligned} \tag{4.2}$$

with Φ is given by

$$\Phi = \begin{cases} \tan\dfrac{\pi\alpha}{2}, \alpha \neq 1; \\ -\dfrac{2}{\pi}\ln|\omega|, \alpha = 1. \end{cases} \tag{4.3}$$

Here, the function $E(\cdot)$ represents the expectation operation with respect to a random variable. α is called the characteristic exponent and indicates the index of stability, while β is identified as the skewness parameter. α and β together determine the shape of the models. Moreover, σ and μ are called scale and shift parameters, respectively. Specifically, if $\alpha = 2$, the α-Stable distribution reduces to a Gaussian distribution.

Usually, it's challenging to prove whether a dataset follows a specific distribution, especially for the α-Stable distribution without a closed-form expression for its PDF. Therefore, when a dataset is said to satisfy a specific distribution, it usually means that the dataset is consistent with this hypothetical distribution and its corresponding properties. In other words, the validation needs to firstly estimate the unknown parameters from a given dataset, and then check the fitting error between the real distribution of the dataset and the estimated one [102].



(a) City A (b) City B (c) City C

FIGURE 4.1: An illustration of the deployment of BSs in three typical cities with geographical landforms.

In this chapter, after an initial examination of data description, we choose different types of heavy-tailed distributions to model the probability distributions of BS density and traffic demand. For comparison, exponential distribution or Poisson distribution are also used for different scenarios. Hereafter, we will present three different cases with real data from cellular networks, spanning from BS deployment to traffic sparsity, and aggregate traffic of MIM.

4.3 BS Density in Cellular Networks

In this section, different from Chapter 3, we try to characterize the BS deployment in one-dimension way. Compare to previous works, stepping back from the spatial pattern modeling where results are not definite, we further examine the BS density distribution in cellular networks. As aforementioned, BSs in urban cities tend to be clusteringly deployed, which implies that the BSs density varies across the whole coverage area. In accordance with the ever-growing heavy traffic of hotspots, the corresponding BS density value can be very high, which makes its distribution to be heavy-tailed. Therefore, in order to characterize this phenomenon, besides the traditional Poisson distribution, we choose several representative heavy-tailed distribution candidates in Table 4.2.

4.3.1 Data Description

In order to reach credible results, we collect a large amount of real data with BSs information from an operator in a well-developed eastern province of China. The collected dataset, containing over 47,000 BSs of GSM cellular networks and serving over 40 million subscribers, encompasses all BS-related records like location information (i.e. longitude, latitude) and BS type (i.e. macrocell or microcell).

TABLE 4.1: The Dataset of BSs and the Related City Information.

Attributes	City A	City B	City C
No. of BSs	8826	5746	4613
City Area	16,847 km^2	9,816 km^2	9,413 km^2
Population	8.844 million	7.639 million	6.038 million
Description	Inland Provincial Capital	Coastal	Coastal

TABLE 4.2: The List of Candidate Distributions and Estimated Parameters in Fig. 4.2.

Distribution	PDF	Estimated Parameters
Generalized Pareto (GP)	$\frac{1}{b}(1+\frac{a}{b}x)^{-(1+\frac{1}{a})}$	a=0.0488, b=3.3502
Weibull	$pqx^{q-1}e^{-px^q}$	p=0.7285, q=0.8279
Log-normal	$\frac{1}{\sqrt{2\pi}nx}e^{-\frac{(\ln x-m)^2}{2n^2}}$	m=-0.1835, n=1.0483
α-Stable	Closed form not always exists. Characteristic Function in Eq. (4.2).	α=0.6207, β=1.0000 σ=0.2053, μ=0.0658
Poisson	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ=1.6759

Based on the coverage area and location information, we divide the dataset into disjoint subsets. Accordingly, we can classify the dataset as subsets of urban areas and rural areas, by matching the geographical landforms with local maps. In this part, for simplicity we primarily take account of urban areas, and try to select the most accurate spatial distribution for BS deployment from various well-known candidate models. Specifically, we choose three typical cities which are capable of reflecting the BS deployment phenomena in metropolis city, big city and medium city, respectively. In Table 4.1, we summarize the detailed information of these selected areas and plot the BS deployment with the geographical landforms in Fig. 4.1, which demonstrates that most BSs are densely clustered while some others are more sparsely deployed.

4.3.2 Fitting and Evaluation

In this section, we conduct the fitting processes to the real data. [68] shows that 10% of the BSs experience roughly about 50-60% of the aggregate traffic load, which implies that the spatial traffic dynamics in cellular networks exhibit heavy-tailed pattern with densely clustering characteristic. Accordingly, in order to fulfill the above nonuniform traffic demand, BSs in urban cities tend to be deployed in clusters as well. Intuitively, the BS density distribution would be heavy-tailed just like the spatial traffic dynamics. Therefore, in order to characterize this realistic phenomenon, besides the traditional Poisson distribution, we choose several popular heavy-tailed candidates in Table 4.2.

Afterwards, based on the large amount of BS location data, we sample one certain city randomly with a fixed sample area size. Then, we compute the spatial density for different 10000 sample

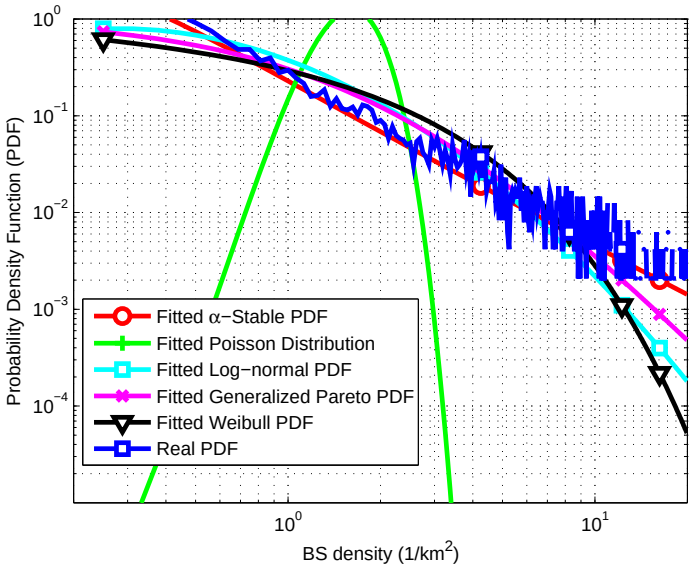


FIGURE 4.2: The log-log comparison between practical BS density distribution in City B with distribution candidates, when sample area size equals 4×4 km^2.

areas and obtain the empirical density distribution, by counting and sorting the number of BSs in each sample area. Next, we estimate the unknown parameters in candidate distributions (except α-Stable distribution) using maximum likelihood estimation (MLE) methodology. For the α-Stable distribution, we estimate the relevant parameters using quantile methods [59], correspondingly build the model to generate the corresponding random variable, and finally compare its induced PDF with the empirical ones.

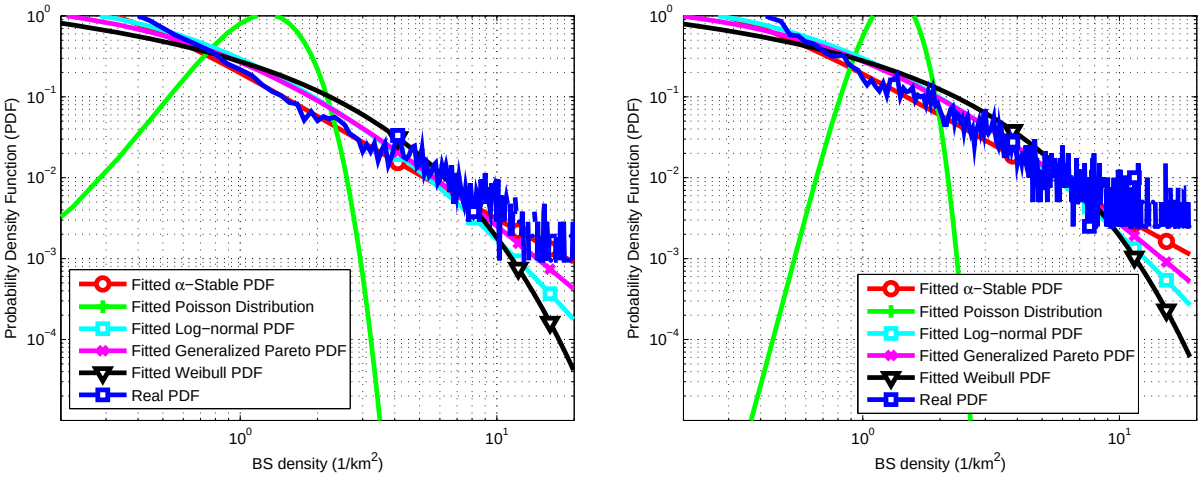
In the first place, we refer to City B as an example, and compute the PDF of BS density under the sample area size 4×4 km^2. After fitting the corresponding PDF to distributions in Table 4.2, we provide the comparison between the empirical distribution with candidate ones in Fig. 4.2. As we can see, the statistical pattern of BS density obviously exhibits heavy-tailed characteristics. Besides, among all candidate distributions, the α-Stable distribution most precisely match the empirical PDF. On the other hand, we provide the numerical comparison in Table 4.3, in terms of RMSE [*]. Indeed, the RMSE results in Table 4.3 show that the α-Stable distribution has the minimum RMSE value (0.0279) while Poisson distribution has the maximum one (0.2537), and once again strengthen this aforementioned conclusion. All of the estimated parameters of the fitted candidate distributions are also listed in Table 4.2.

Meanwhile, to verify the general accuracy of candidate distributions, we change the sample area sizes to 3×3 and 5×5 km^2 respectively, and plot the related results in Fig. 4.3(a) and Fig. 4.3(b). Obviously, compared to other candidate distributions, the α-Stable distribution still provides the most accurate fitting results for the BS density distribution in City B.

[*]Root mean square error.

TABLE 4.3: RMSE Values after Fitting Candidate Distributions to Empirical One in Three Cities.

City	Sample Area Size (km^2)	α-Stable	Poisson	Lognormal	GP	Weibull
A	3×3	0.0105	0.1214	0.0207	0.0274	0.0361
	4×4	0.0177	0.1465	0.0269	0.0339	0.0418
	5×5	0.0286	0.1702	0.0293	0.0357	0.0432
B	3×3	0.0207	0.2088	0.0658	0.0770	0.0924
	4×4	0.0279	0.2537	0.0905	0.1017	0.1151
	5×5	0.0300	0.2913	0.0971	0.1085	0.1217
C	3×3	0.0373	0.2332	0.0513	0.0755	0.0910
	4×4	0.0451	0.2918	0.0697	0.0948	0.1076
	5×5	0.0487	0.3405	0.0705	0.0960	0.1064



(a) 3×3 km^2 (b) 5×5 km^2

FIGURE 4.3: The results after fitting candidate distributions to BS density in City B, when sample area varies.

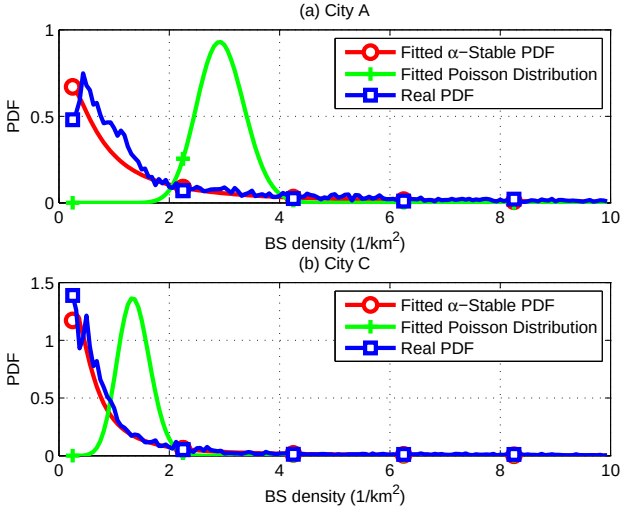


FIGURE 4.4: The comparison between BS density distribution and α-Stable distribution in City A and City C, when sample area size equals 4×4 km^2.

In order to examine the geographical impact on the fitting results, we further analyze the density distribution of BSs in City A and City C using a sample area size of 4×4 km^2. Due to the factor of geographical irregularity, there is a noticeable gap between the α-Stable distribution and the empirical PDF of Cities A and C in comparison to City B. Nevertheless, as shown in Table 4.3 and Fig. 4.4, it can be observed that, the α-Stable distribution could match the practical one in both cities, with RMSE values equaling 0.0177 and 0.0451 respectively and being less than those of other candidate distributions. Moreover, the same conclusions goes with sample area sizes of 3×3 and 5×5 km^2, as testified in Table 4.3.

Based on the extensive analyses above, we could confidently reach the following remark.

Remark 4.1. The spatial pattern of deployed BSs exhibits strong heavy-tailed characteristics. Based on the large-scale identification, the α-Stable distribution manifests itself as the most precise one. On the contrary, the popular Poisson distribution appears to be an inappropriate model for the BS density distribution, in terms of the RMSE.

4.3.3 Conclusion

In this section, based on the real BS deployment information of on-operating cellular networks, we carry out a thorough investigation over the statistical pattern of BS density. Our studies show that the distribution of BS density exhibits strong heavy-tailed characteristics and the widely adopted Poisson distribution severely diverges from the realistic distribution. Instead, the α-Stable distribution, which is also found in the traffic dynamics of broadband networks, can most precisely match the BSs deployment in cellular networks.

4.4 Spatial Density of User Traffic Demands

Besides BSs, the spatial distribution of traffic demand is also very important for characterizing the operating performance of cellular networks. In this section, we try to analyze the spatial density of data traffic based on real measurements from a large number of BSs, where the aggregate traffic are collected.

After that, we try to uncover the statistical relationship between BSs deployment and traffic demand distribution which is highly coupled with each other. For example, operators need to deploy new transmit tower for a newly built residential area. Therefore, after the separate modeling of BS density and traffic density, we conduct the hypothesis test for the linear relationship between these two variables.

4.4.1 Data Description

The dataset, collected from two kinds of networks (i.e., 2G and 3G cellular networks), includes traffic and BSs information of City A and City B. The data traffic is measured in the unit of bytes that each BS transmits to the covered users in a one-hour interval.

Specifically, we convert the longitude and latitude values of each BS to distance coordinates, and plot the actual geographic location on a 2D plane. Meanwhile, according to the city structure, four sample regions named Urban1, Rural1, Urban2 and Rural2 are selected. Obviously, BSs density in rural areas is far smaller than that in urban areas. Moreover, most BSs in rural areas exhibit spatial sparsity while BSs are aggregated densely in urban area.

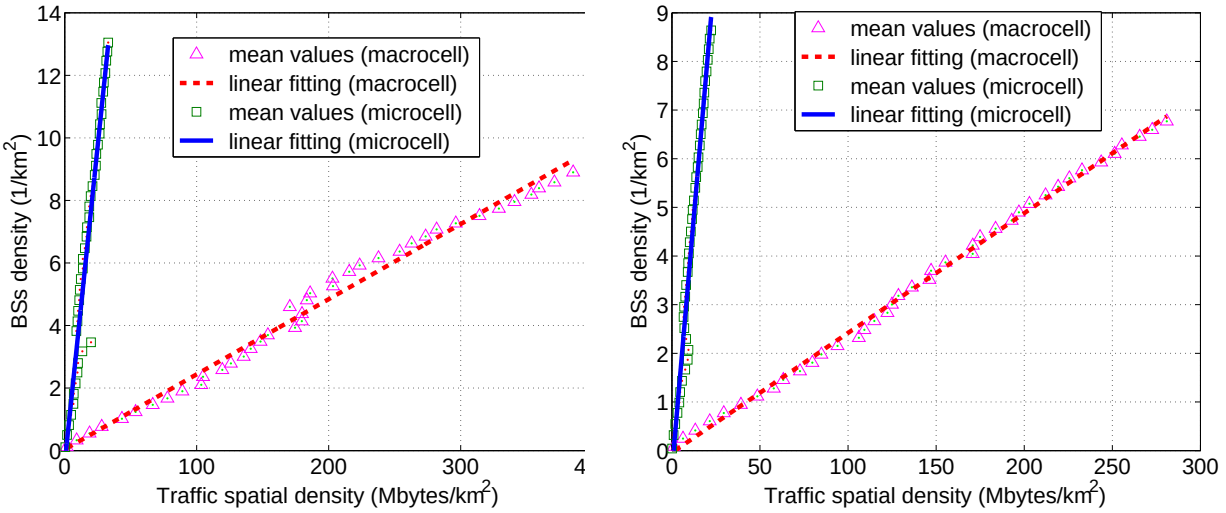
4.4.2 Spatial Distribution of Traffic Demand

Humans with similar social behaviours tend to live together, which leads to various traffic hotspots and causes BSs to be deployed as clusters in the corresponding areas. [68] pointed out that less than 10% of the subscribers generate 90% of the traffic load while 10% of the BSs carry 50%-60 % of the traffic load, which demonstrates significant traffic imbalance and BSs inhomogeneity in cellular networks. Hence, based on the dataset described above, we aim to reveal the inhomogeneity of BSs and traffic distributions.

Firstly, a square sampling window with size S, is selected randomly. Then, we compute the number of BSs (N_{BS}) within this window and their aggregate data traffic (V_{TR}). Thus, one tuple (N_{BS}, V_{TR}) is recorded for each sampling experiment, and the same procedure is repeated 10000 times to obtain enough tuple records. Accordingly, BSs density (λ_{BS}) and traffic spatial density (λ_{TR}) are identified as follows:

$$\begin{aligned} \lambda_{BS} &= \frac{N_{BS}}{S}, \\ \lambda_{TR} &= \frac{V_{TR}}{S}. \end{aligned} \tag{4.4}$$

Considering the real situations where heavy-tailed phenomenon does exist in BSs and traffic spatial distributions, we take the α-Stable distribution as the fitting candidate. Afterwards, we use the α-Stable model, produced by the aforementioned estimated parameters, to generate some random variables, and compare the induced PDF with the empirical one. Therefore, after fitting an α-Stable distribution to BSs density and traffic spatial density in City A (sampling window size is 3×3 km^2), they both better obey the α-Stable distributions obviously (similar to the findings in [24], [113]). For City B, the α-Stable distribution is also applicable.



(a) 5×5 km^2 (b) 7×7 km^2

FIGURE 4.5: The fitting results of Urban1, when sampling window size varies.

4.4.3 Linear Dependence Between BSs and Traffic

Geographic limitations, as well as city structures, lead to the diversities of population and traffic demand in different regions. Accordingly, the mobile operators adapt their BSs to where the subscribers generate the most traffic. In other words, BSs density is closely related to traffic spatial density. In this section, we will check whether there is any intrinsic correlation between these two quantities.

To ease illustration, Urban1 is taken as a representative example. With the sampling window size being 5×5 km^2, fitting results are depicted in Fig. 4.5(a). Evidently, BSs density and traffic spatial density exhibit strong linearity regardless of the BS type. Besides the visual observation, R-square (R^2) value is also adopted as a performance metric to evaluate the goodness of fit. The closer R^2 value to 1, the better is the fit. From Table 4.4, the R^2 value of macrocell and microcell equals 0.9890 and 0.9503, respectively. Therefore, linear model is reasonable to characterize the spatial correlation between BSs deployment and traffic spatial distribution, which can be stated as follows:

$$\lambda_{\mathrm{BS}} = k\lambda_{\mathrm{TR}} + t. \tag{4.5}$$

Here, k is the slope value that represents the needed number of BSs per unit of spatial traffic.

To further verify the general accuracy of linear model, sampling window size of 7×7 km^2 are similarly studied, and the corresponding results are illustrated in Fig. 4.5(b). Clearly, the sampling window size variation does not violate the linearity. Meanwhile, same tests are carried out in Rural1 and similar conclusions are derived but with different fitting parameters. More detailed numerical results are displayed in Table 4.4.

TABLE 4.4: Fitting Parameters of Different Geographic Scenarios.

		Urban1		Rural1	
BS type Window Size (km^2)	Sampling k	R^2	k	R^2	
macrocell	3×3	0.0226	0.9609	0.0258	0.9887
	5×5	0.0241	0.9890	0.0261	0.9970
	7×7	0.0246	0.9977	0.0262	0.9956
microcell	3×3	0.3916	0.9245	0.3161	0.8357
	5×5	0.4029	0.9503	0.2919	0.8739
	7×7	0.4196	0.9638	0.3060	0.8737

On one hand, linear model keeps better fitting performance no matter the sample region is (urban or rural). On the other hand, the key parameter slope k is closely associated with the BS type, without dependence on the sampling window size. These findings indicate that BSs deployment is deeply influenced by subscribers' demand as well as the corresponding traffic dynamics over the space, and imply that BSs density and traffic spatial density have almost identical heterogeneity feature. Interestingly, it is consistent with previous findings that both BSs deployment and traffic spatial pattern demonstrate the same characteristics (i.e., obeying the α-Stable distribution).

In practice, with the increase of the traffic load, it is impossible for the number of BSs to grow linearly and infinitely, due to the physical and performance constraints of cellular networks. Consequently, there should be a certain critical state where the available service capability is pre-determined, and if the traffic demand increases continuously, there would be a network evolution (i.e., upgrading from 2G to 3G, then 4G). In that regard, an explanatory outline about how cellular network architecture evolves is illustrated with different slope k in Fig. 4.6. Surely, the performance improvement of networks expects to serve more traffic demand with smaller number of BSs.

Based on the extensive analyses above, we can reach the following remark.

Remark 4.2. The BSs deployment and traffic distribution exhibit a strong linear dependence, which suggests that the heterogeneity feature of the two quantities is almost identical. The slope k in the linear model implies the capacity performance of specific BSs and can be adopted as a valuable performance metric to evaluate the long-term evolution of cellular networks.

4.5 Temporal Characterization of Mobile Instant Message

In this section, we turn to the temporal dimension of traffic demand in cellular networks, and takes MIM traffic as an example. As described in Chapter 2, the traffic demand also exhibits temporal clustering nature as spatial dimension. Therefore, after the preliminary description of

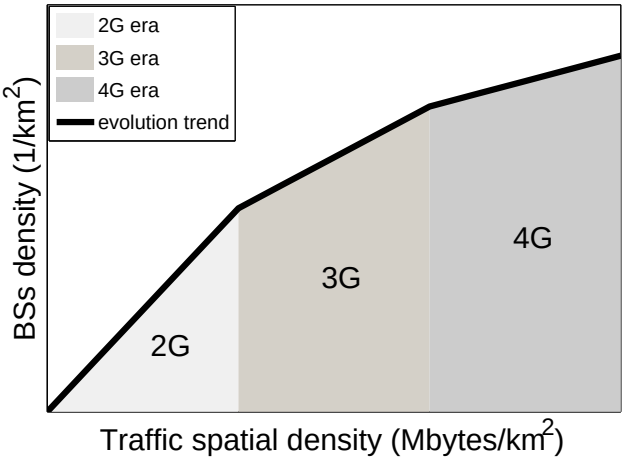


FIGURE 4.6: The cellular network evolution trend on capacity based on BS-traffic linearity [53].

data measurements, we choose several heavy-tailed distribution to conduct the fitting process, in order to find the most appropriate model for temporal distribution of traffic demand.

4.5.1 Data Description

In order to build primary models, we collect measurements of the MIM traffic from the on-operating cellular networks. Our datasets collected from the Gb and Gn interfaces [111], covering about 15000 GSM and UMTS BSs in an eastern provincial capital within a region of 3000 km^2, could be classified into two categories in terms of the corresponding resolutions (i.e., IML traffic and aggregated traffic). The 1-month measurement records of IML (Individual Message Level) traffic are collected from 7 million subscribers, and contain timestamps, cell IDs, anonymous subscriber IDs, message lengths, and message types. In contrast, the measurement records of aggregate traffic possess coarser resolution than those of IML traffic, and merely specify per 5-minute traffic volume of roughly 6000 BSs in the same city on September 9th, 2014. Fig. 4.7 plots the snapshots of the aggregate traffic at three different moments in a region.

4.5.2 Fitting and Evaluation

In this part, we conduct the fitting process to the real data collected on two different levels, i.e., IML traffic and aggregate traffic.

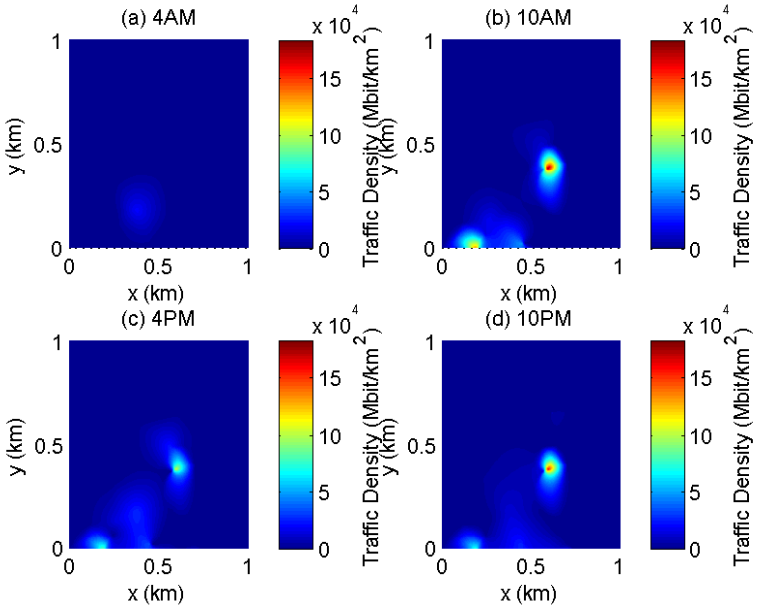


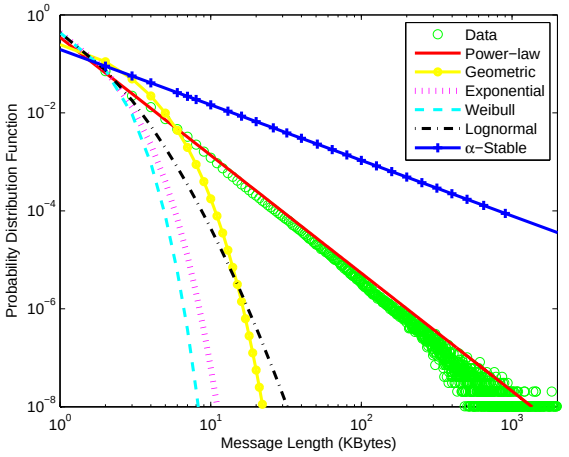
FIGURE 4.7: The snapshots of aggregated traffic at three different moments in a region containing 23 BSs.

IML Traffic on User Level

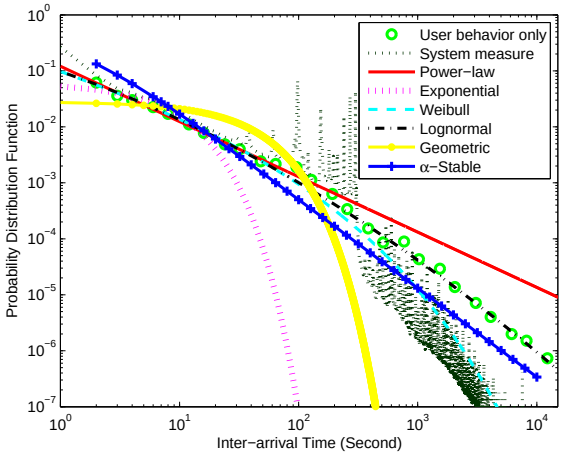
In order to understand the IML traffic nature of MIM services, we firstly calculate the PDF of message length, and then fit them to common heavy-tailed distributions. Specifically, during the fitting procedures, we obtain the unknown parameters in candidate distribution functions (except α-Stable models) using MLE method. For α-Stable models, we estimate the relevant parameters using quantile methods [60], correspondingly build the models to generate some random variable, and finally compare its induced PDF with the exact (empirical) one.

In Fig. 4.8(a), we provide the corresponding results after fitting candidate distribution functions to the empirical PDF of message length. Interestingly, Fig. 4.8(a) demonstrates that instead of geometric distribution function recommended by 3GPP [1], power-law distribution (i.e., $0.347x^{-2.407}$) could most accurately approximate the empirical PDF of message length. Furthermore, the RMSE is also applied to quantitatively find the fittest distribution function. The results also show that the PDF of message length is most appropriately to be modeled by a power-law distribution.

On the other hand, according to the time-stamps of messages, we calculate inter-arrival time t between consecutive messages in the order of second, and examine the fitting preciseness of MLE estimated candidate distribution functions to the corresponding PDF. Fig. 4.8(b) depicts the related fitting results compared to the empirical data (see legend: User behavior only) with the inter-arrival time from 2nd to 3000th second. Compared to the exponential distribution function recommended in [1], Fig. 4.8(b) shows that lognormal distribution function



(a) Fitting results of the length of MIM messages.



(b) Fitting results of the inter-arrival time of MIM messages.

FIGURE 4.8: Fitting results of MIM activities' message length, inter-arrival time.

(i.e., $\frac{1}{\sqrt{2\pi}\times 2.975x}e^{-\frac{(\ln x-2.36)^2}{2\times 2.975^2}}$) exhibits superior fitting preciseness for the inter-arrival time of MIM messages.

Remark 4.3. Compared to the geometric and exponential distribution functions recommended by 3GPP [1], power-law and lognormal distribution functions are more suitable to model the statistical pattern of message lengths and inter-arrival time of consecutive messages, respectively.

Aggregated Traffic on BS Level

In this part, from the perspective of one whole BS, we examine the fitting results of aggregate traffic within one BS to candidate distributions. Fig. 4.9 presents the corresponding PDF comparison between the simulated results and the real aggregate traffic in one randomly selected BS. By taking advantage of a similar methodology, Fig. 4.9 implies that the traffic records in these selected areas could be better simulated by α-Stable models. Similarly, it shows that α-Stable models lead to better fitting accuracy in terms of RMSE. Furthermore, Fig. 4.10(a)$\sim$(d) verify the fitting preciseness of empirical data to α-Stable models in another four randomly selected BSs, and the curve in Fig. 4.10(e) indicates the low level of fitting error.

On one hand, the universal existence of α-Stable models implies and contributes to understand the intrinsic self-similarity feature in MIM traffic [24]. On the other hand, the reasons that MIM traffic universally obeys α-Stable models can be explained as follows. Previous work unveils that the length of one individual MIM message follows a power-law distribution. Meanwhile, the distribution of aggregate traffic within one BS can be regarded as the accumulation of lots of IM messages from diverse users. Therefore, according to the generalized central limit theorem, the sum of a number of random variables with power-law distributions decreasing as $|x|^{-\alpha-1}$

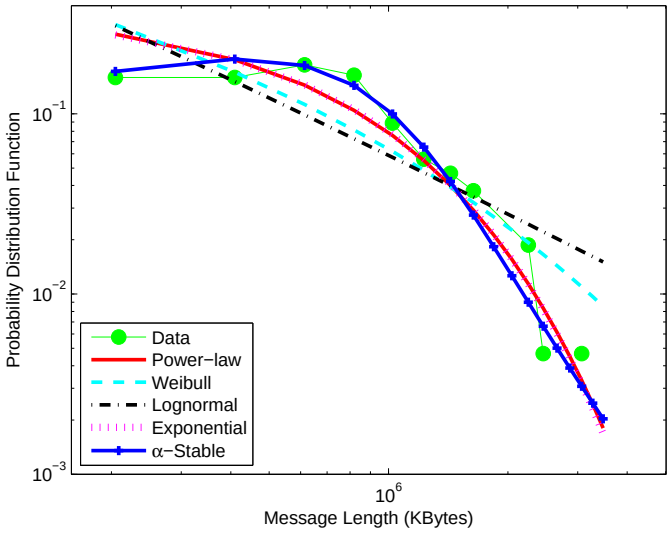


FIGURE 4.9: Fitting results of candidate distributions to empirical aggregated traffic in one randomly selected BS.

where $0 < \alpha < 2$ (and therefore having infinite variance) will tend to be an α-Stable model as the number of summands grows. Interestingly, Fig. 4.10(f) shows that the PDF of parameter α obtained by fitting aggregated traffic in different cells to α-Stable models, and reflects the fitting values of α mostly fall between 1.136 to 1.515, while the slope of power-law distribution for IML traffic is 2.407. These fitting results prove to be consistent with the theory from the generalized central limit theorem [47].

Remark 4.4. The aggregated traffic within one BS, following α-Stable models, can be explained as the accumulation of a number of power-law distributed messages.

Conclusively, we investigated the traffic characteristics of MIM services from two different viewpoints. For IML traffic, we showed that message length and inter-arrival time better follow power-law and log-normal distribution, which are quite different from the recommendation by 3GPP. For aggregate traffic within one BS, we revealed the accuracy of applying α-Stable models to characterize this statistical pattern, and extended the suitability of α-Stable models for traffic in both fixed core networks and cellular access networks. Besides, following the generalized central limit theorem, we built up the theoretical relationship between distributions of IML and aggregated traffic. These heavy-tailed traffic models of MIM service could contribute to the design of more efficient algorithms for resource allocation and network management in cellular networks.

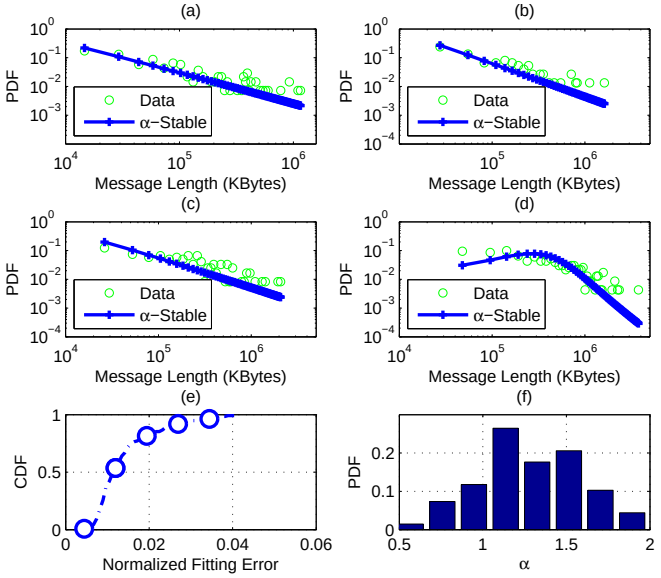


FIGURE 4.10: (a)∼(d): Fitting results of α-stable models to empirical aggregated traffic in another two randomly selected BSs; (e): The preciseness error CDF (cumulative distribution function) for all the cells after fitting; (f): The PDF of α estimated for aggregated traffic in different cells.

4.6 Conclusion and Discussion

In this chapter, we tried to characterize the statistical distribution of BS deployment and traffic demand, in both spatial and temporal domains.

Firstly, we investigated the spatial density of BSs, in order to characterize the variation of BS deployment across the whole cellular networks. In detail, after revealing the spatial clustering nature of BS deployment, we chose several heavy-tailed distributions to be fitted to the real data along with the traditional exponential distribution. Based on the RMSE, we found that the α-Stable distribution is the most accurate model for the BS density, which differs from the traditional assumption that BSs tend to be deployed independently.

Secondly, we turned to the spatial distribution of traffic demand in cellular networks. Similar with the case of BSs deployment, the spatial density of traffic volume on cell level also exhibits heavy-tailed property. After fitting process and performance evaluations, we again revealed the accuracy of the α-Stable distribution for this scenario.

Thirdly, due to the same features of spatial density for BSs deployment and traffic demand, we investigated the quantitative relationship between those two variable. Based on the same data set, we found out that there is a kind of linear dependence between BS and traffic density. That's to say, where there are more BSs, the traffic demand tends to increase, or vice versa.

Indeed, the slope between these two quantities are found to be constant related to the technology development, which is quite reasonable.

Besides the spatial analysis, we also conduct the temporal analysis of cellular traffic, and taking the popular MIM traffic as an example. For each sampled BS, we conduct the traffic density characterization on the time scale, and again find that the α-Stable distribution is most consisted with the collected traffic volume.

4.6.1 Connecting the Dots

After introducing the fitting results on different dimensions, we come to wonder why the α-Stable distribution is the most accurate model in different cases. Is there any internal causes for these results? Although the clustering nature is revealed in different dimensions, there are different statistical distributions for characterizing the densities. Within those candidate models, the α-Stable distribution is the most general one, which covers the Gaussian distribution, Levy distribution and Cauchy distribution as special cases with specific parameters.

Indeed, like Gaussian distribution, the α-Stable distribution can be derived from a so-called generalized central limiting theorem where the individual variables follow a variance unlimited power-law distribution. In our cases, the traffic volume of a specific BS is aggregated from many mobile users, whose traffic usage pattern diversifies. Therefore, if the user traffic volume on time scale can be verified to be power-law distributed, then the α-Stable distribution of aggregate traffic volume is reasonable.

In the spatial case, the BS density exhibits heavy-tailed phenomenon, where a small part of regions covers most of the BSs. Besides, the traffic density for different cells also tends to be highly skewed, where dense urban areas have very high traffic density while rural areas are shown to be less loaded. To explain that, we first assume that the spatial density of BS deployment and traffic demand are linear dependent, which is quite reasonable considering the deployment strategy of telecommunication operators. Therefore, the problem why the α-Stable distribution fits for the traffic density and BS density reduce to the generation of aggregate traffic in single BS. Actually, the traffic volume of a BS is generated by lots of users within the coverage area. As we introduced in Chapter 2, the traffic usage patterns of mobile users vary across the population, thus the variance of the individual traffic records can be assumed to be large enough. According to the generalized central limiting theorem, this assumption comes to the α-Stable distribution of the aggregate traffic volume on BS level.

Therefore, from our analysis, the α-Stable distribution for BS density and traffic density are rooted in the individual traffic usage of mobile users, which need to be assumed as power-law distribution.

Chapter 5

Cooperative Caching and Dynamic Multicast: Making Cluster Benefit

Contents

In the previous two chapters, through the analysis of real data, we not only obtained qualitative judgments on the clustering nature of cellular networks, but also gave quantitative descriptions of them on different dimensions. Intuitively, in this chapter, we will explore what change can

such an aggregation feature (space, time or content) bring about on the service performance of cellular networks. Compared to the traditional uniform distribution assumption, will this change degrade the overall performance or bring opportunities for improvement? In the following sections, we will propose two efficient strategies based on practical service scenarios, so as to make a rational use of the clustering effect in different dimensions.

5.1 Introduction

5.1.1 Background

Traditional cellular networks have evolved from the first generation of analog communications to the current fourth generation of digital communications. Iteratively enhanced physical layer technologies have greatly increased network capacities, such as different multiple access technologies, modulation and demodulation methods. According to the Shannon's theory, the technical gains brought by the physical layer gradually become saturated, which cannot match the rapid increase of user traffic demand in the current mobile Internet era. From another point of view, the traditional technologies are focused mainly on improving the service capacity of the network itself, but not much on the service target of the network (mobile users or traffic demand). If we can provide a flexible service strategy based on different demand characteristics from the perspective of satisfying user needs, the existing physical layer technologies may provide overall service capabilities that could not be achieved before.

Therefore, the focus of our attention should shift from the underlying access transmission technology to higher-level user demand descriptions and service capabilities. Specifically, such a general idea can be decomposed into two paths. On one hand, the measured data is used to understand and characterize the user's real demand. On the other hand, given those real-data-oriented demand characteristics, how the underlying technology should be fine-tuned or orchestrated to achieve greater service capacity. In the previous chapters, we basically completed a spatio-temporal analysis of the infrastructure and traffic demand in the cellular network through a large amount of measured data. Coupled with the existing content popularity analysis in literature, we can presumably describe the clustering nature of a general mobile network. Next, we mainly consider what kind of service mechanism should be adopted to maximize the capacity gain that clustering feature may bring.

5.1.2 Related Works

The traditional cellular network service flow complies with a request-and-serve mechanism. That is, the user firstly initiates a content request to the connected BS , and the BS fulfills the

user request by coordinating with other network entities. In this way, each user request needs to go through a complete link, for example, from the user end through the BS to the access network, and then access the public telephone network or the external Internet through the core network. Such a service strategy can indeed provide a reliable quality of service for each user request, but at the same time it also exposes problems of insufficient usage of communication resources, especially in the case of heavy traffic and significant content homogeneity. Considering the aggregation characteristics of actual user requests, caching and broadcasting technologies have been extensively studied in cellular networks in recent years [69], which utilize the spatial and temporal aggregation characteristics of user data requests separately, although there is not complete formula description on the logic behind them.

In actual user data requests, related statistics [30, 70] found that most user requests in the cellular networks direct to a small amount of popular contents, which provides a reasonable scenario for the application of caching technologies. Caching popular contents on the CN [*] gateways, BSs or even mobile devices can greatly reduce the capacity requirements of peak traffic on the cellular backbones, and can also effectively reduce the average waiting time of mobile users. According to the cache location of content, we can divide the mobile cache technology into several categories, which are the CN cache, the RAN cache, and the mobile device cache [94].

By caching popular contents in CN, the user request can save the back and forth journey of obtaining content from the Internet. This aspect reduces the traffic flow from the CN gateway to the external Internet, and on the other hand, it can also partially shorten the user's content access delay [98]. At the same time, the network elements in CN can also cache enough popular contents to further increase the probability of hitting cached content. The caching in RAN is closer to mobile users than core network caching, thus more attractive in terms of delay reduction and backbone traffic offloading, but it is not as better in storage capacities [2]. On the other hand, since the number of mobile users under a single BS is limited, the RAN caching cannot clearly form the homogeneity of user demand content. Finally, researchers began to consider directly caching popular contents on mobile devices [45]. In this way, mobile users supporting D2D (device to device) transmissions can exchange cached contents themselves, and do not need too much intervention from the cellular networks. Such a buffering method can greatly ease traffic congestion in the backbone, and can also reserve more resources for other transmission tasks while reducing the user delays. However, the D2D transmission method under the caching strategy also has great limitations. The first is that the storage space of mobile devices is very small and cannot cache enough contents. Therefore, it is necessary to introduce cooperative transmissions between different devices, which increases the complexity of protocol design. Secondly, the frequency band of D2D transmission, the potential interference and the security issues remain as open problems.

[*] Core networks.

In addition to caching technology, broadcast or multicast technology is also considered to improve the transmission efficiency in cellular networks [6]. Multicast, which is a one-to-many transmission strategy, is suitable for a service scenario in which multiple users request the same content at the same time (or within a short interval). In such a case, the BS can send the same content to many users within its coverage area, to achieve the purpose of once-transmission and multiple-receptions. Such a strategy can not only increase the capacity of the entire access networks, but also effectively reduce the energy consumption of the BS. However, it works at the cost of increasing user delay, because the content request needs to wait for the multicast until the number of requests achieved a pre-defined threshold. In general, the multicast technologies in cellular networks utilize the characteristics that multiple geographically similar users obtaining the same content in adjacent time, which is essentially based on the spatial and temporal clustering of mobile users' traffic demand. In practice, multicast or broadcast can be combined with other promising technologies to further exploit the characteristics of the cellular network for potential capacity improvement.

For example, multicast technology can be combined with D2D short-distance transmissions [61]. Among all the mobile users requesting the same content, some receiving users having better channel quality as the BS perform multicasting can serve as relays. After these relay users obtain the required content from the BS, they spread the content to all the requesting users through D2D transmission. Such a combined multicast and D2D service approach can not only exert the advantages of the high efficiency of multicast transmissions, but also makes full use of the differences in downlink channel conditions of mobile users.

In addition, the combination of multicast technology and BS cooperation has also been investigated. Generally, in current 4G networks or future 5G networks, heterogeneity of BSs is an important attribute: macro BSs have better coverage performance due to lower spectrum, and micro BSs have narrow coverage and high transmission rates with higher frequency. Such a functional difference also enables the macro BS and the micro BS to play different roles in the multicast scenario. Specifically, according to the user's content request, multicast can be initiated within the coverage of the macro BS, so that the advantage of a large number of user request bases can be fully utilized. After the multicast is begun, the user requests initiated after the cutoff time can also be added to the queue to participate in receiving the remaining transmission content. After that, the content data packet that has been multicasted before can be passed through the micro BS with the high rates. In this way, the original coarse-grained multicast method can be decomposed into two steps, where most of the requested content can be multicasted through a macro BS with a wide coverage and the remaining parts of the content can be transmitted through a selected micro BS according to the corresponding channel condition. Compared to the previously mentioned D2D transmission, the BS cooperation-based multicast technology does not require the mobile users to reserve the storage space for the relay transmission, and also has a higher energy efficiency.

In addition, because different users who initiate the same content request have different downlink channel conditions, the network coding is often combined with multicast technology to meet users' diverse rate requirements [88]. Simulations showed that such an hybrid strategy can achieve a throughput improvement of 30% to 45% based on the use of multicast technology alone.

In summary, both caching and multicast technology use the clustering nature of content requests of the cellular networks in the time, space and content domains, to improve the overall service efficiency, such as reducing access delay, increase spectrum efficiency, reduce BS power consumption. Although related works recognized the potential impact of user request patterns or network facility placement on overall service efficiency, and proposed corresponding policies (cache or multicast) analyzes their respective performance advantages. However, the user request mode or BS distribution model adopted by most works cannot accurately reflect the reality. For example, in the RAN caching strategy, the popular contents are stored on the BS, and the mobile users requesting the corresponding content select an appropriate BS according to the principle of proximity for content acquisition, thus the BS locations tend to be very important in this scenario. However, the spatial locations of BSs are usually determined as a uniform distribution, which is not consisted with our previous conclusion that BSs tend to be clusteringly distributed. In fact, the aggregated distributed BSs provide a possibility for more efficient inter-cluster cooperation, and such an approach can hopefully improves the overall efficiency of the caching technology. On the other hand, in the performance analysis of multicast technologies, the time arrival mode of user content requests is usually assumed to be a uniform Poisson process (for mathematically derivable properties), which is also inconsistent with the bursty nature claim we made in Chapter 4.

5.1.3 Approach and Contributions

In the above discussions, we summarized the inherent logic behind these potential technologies, and pointed out possible ways to improve. As mentioned before, unlike traditional uniform distributions, the burstiness of content requests and the spatial clustering of BSs deployment in cellular networks may provide a practical perspective for these two technologies. Therefore, the purpose of this chapter is to investigate how the aggregation of user requests and BSs can bring about performance gains for cache and multicast strategies.

For comparison, we first need to calculate the performance gain that caching and multicast policies can achieve under the uniform distribution of content requests and BSs deployment. Secondly, we calculate the corresponding performance gains under the aggregately distributed content requests and BSs. Further, we try to adjust the traditional caching and multicast strategies to be consisted with the clustering nature oriented from the real data measurements.

For example, according to the spatially clustered distribution of BSs, the cooperation within BS clusters can be introduced to the traditional caching strategies. In this way, mobile users can access more cache contents on the BSs, thereby significantly reducing content acquisition time. For another example, the arrival rates and clustering degrees of different content requests can be predicted based on the real-time records, based on which the corresponding multicast threshold would be dynamically adjusted so as to improve the overall delay and power consumption performance.

In conclusion, we proposed a cooperative caching strategy in RAN based on the spatial aggregation of BSs and a dynamic unicast/multicast strategy based on the temporal aggregation of content requests in Chapter. According to the theoretical and simulation results, we found that the proposed 'Caching as a Cluster' strategy can significantly reduce the average latency of users especially in the inhomogeneous BS deployment scenario, and the dynamic unicast/multicast strategy can not only reduce the average latency of content requests but also diminishing the average power consumption of BSs especially under the bursty request arrival patterns.

5.2 Clustering-based Cooperative Caching in Cellular Networks

This section will discuss the probabilistic caching strategies in spatially clustered cellular networks. Thanks to the content preference of mobile users, proactive caching can be adopted as a promising technic to diminish the backhaul traffic and to decrease the content delivery latency. However, basically there are two obstacles to accomplish the caching policy, i.e., the limited storage capacity of small cells to cache a large amount of multimedia contents, and the too small number of users under each BS to imply the content aggregation effect. Traditional caching strategies of the BS only concern its local requests from the connected users through wireless links, but neglects the potential benefit from the cluster feature of the network infrastructure and user traffic demand. In this section, we proposed a new policy called 'Caching as a Cluster', where small cells can exchange contents with each other to fulfill every user request within the cluster of BSs. Intuitively, this cooperation between BSs makes a difference to decrease the content delivery latency of mobile users in clustered cellular networks as testified in our numerical simulation [114].

5.2.1 Motivation and Objectives

In recent years, proactive caching has been widely investigated in cellular networks [9], which is motivated by the content preference of mobile user requests, to increases the network capacity and reduces the delivery latency. Mostly, researchers consider the BSs as storage anchors [57] or even the user equipments in D2D scenarios [45]. However, there are two intrinsic obstacles to

accomplish the caching policy in the wireless part of cellular networks, i.e., the limited storage capacity of single BS to cache large amount of multimedia contents, and the too small number of users under each BS to imply the content aggregation effect [94]. To solve those problems, we propose a probabilistic caching scheme in spatially clustered cellular networks, where all SBSs (small BSs) within a cluster can exchange stored contents cooperatively to fulfill different content requests which can be aggregated together to provide more apparent content preference phenomenon.

Recently, many researches focus on the small cell caching, such as the FemtoCaching proposed in [84], where the authors consider that UEs can access to several different SBSs. Similarly in [10], it's declared that the most popular content caching policy is not the optimal under the high coverage regime. In general, all these related works assume that UEs directly connect to the SBS which stores the demanding content [16, 52]. However, this kind of procedure actually increases the interference in wireless environment, thus decreasing the overall throughput of the whole cellular networks.

Besides, there is a tendency to centralize the base band units of nearby SBSs with fiber-based link, plenty of wired bandwidth would be available for the caching content sharing among the cluster of SBSs [19]. From this point of view, we can reformulate the content placing problem in wireless caching scenario by adding the content sharing cooperation between SBSs, under which the interference can be mitigated [94].

Our research here differs from related works in two aspects. First, we consider probabilistic caching strategies rather than completely deterministic content placement methods. The traditional deterministic content caching problem is usually NP-hard, and there are no effective high-efficiency solutions, and the probabilistic caching problem is simpler to solve since variables are all continuous values, while achieving the same performance as deterministic content placement strategies when the number of content is large enough. Secondly, unlike the independent serving process in traditional caching strategy, we propose a new idea of cooperative caching in the aggregated distributed SBS scenario which more or less relaxes those two aforementioned obstacles.

Using the transmission bandwidth between the SBSs within the cluster, we have shown through simulation analysis that the proposed cooperative caching strategy outperforms the popular greedy one and the uniform one in terms of the average delivery latency after considering the queue waiting time. This section is organized as follows: after introducing the system model and formula description for the corresponding problem, the solution is proposed along with the theoretical analysis and the numerical results. Afterwards, the conclusion and potential improvement are given at last.

5.2.2 Model Description and Problem Formulation

In this section, we introduce the system model and formulate the probabilistic caching strategy into a mathematical minimization problem.

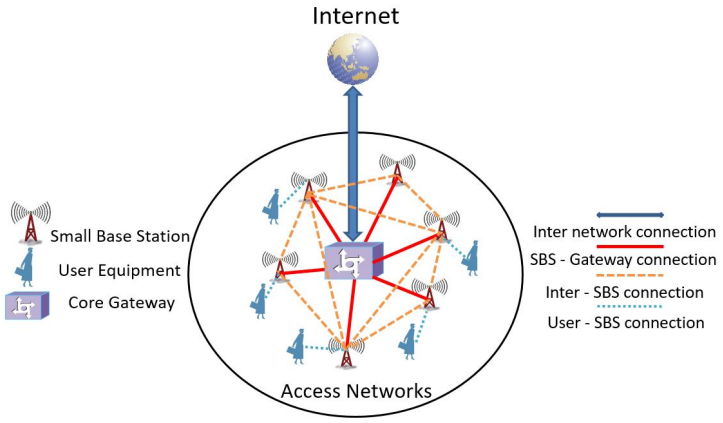


FIGURE 5.1: Cooperative content caching in clustered cellular networks.

Here, we consider a spatially clustered cellular network, which consists of a cluster of SBSs and their connected core gateway, as illustrated in Fig. 5.1. We assume that the MBS takes responsibility of the control plane, which orchestrates the placement and offloading of cached contents, thus not displayed in this data plane plot.

Utilizing the content preference of mobile users, SBSs can cache popular contents in their local storage to reduce the overall latency. Different from related works, here we assume that within the SBS cluster, cached contents can be exchanged between all SBSs to serve each user request. In detail, UE sends data requests to its connected SBS, then the SBS will firstly check whether the requested content is in its storage (local hit) or not. If not, it will offload this request to another available SBS according to a local reference table which is managed by the MBS. Then this requested content can be transferred to the UE through the inter-link between these two SBSs (cluster hit). Otherwise, the connected SBS should turn to the gateway which is assumed to contain the whole library (global hit). Generally, the delivery latencies are distinctive for different hit scenarios and routing paths, which is the basic principle here to design the caching strategy.

In detail, we assume that the storage capacity of each SBS is C_s and there are N_s equivalent SBSs within each cluster, where the bandwidth between each pair of SBSs is B. Additionally, the content requests rate under each SBS is λ, and the serving rate of each SBS is μ, with $\lambda < \mu$. Because of the requests offloading, the arriving rate for each SBS depends on the caching strategy, thus we adopt the queuing-based average waiting time for local and cluster hits, instead of constant latency. Furthermore, the global hit latency D_{gh} is assumed to be constant and significantly larger than those of local and cluster hits (D_{lh} and D_{ch}), due to

TABLE 5.1: Probabilistic caching parameters description.

Parameter	Description	Typical Value
λ	Requests rate under each SBS	100 s^{-1}
μ	Serving rate of each SBS	250 s^{-1}
N_s	Number of SBS in the cluster	10
C_s	Storage capacity of SBS (contents count)	20
B	Bandwidth between each pair of SBSs	800 Mbps
D_{gh}	Delivery latency from core gateway	30 ms
N_c	Number of different contents	100
Z	Size of each content	50 MB
γ	Skewness of Zipf's law	0.65

the long transmission distance and larger serving capacity [104]. Above all, the imbalanced distribution of requested contents is represented as Zipf's distribution (assume each content has the same size $Z = 50MB$), then the probability of the ith most popular content is calculated as:

$$f_i = \frac{i^{-\gamma}}{\sum_{j=1}^{N_c} j^{-\gamma}}, \tag{5.1}$$

where N_c is the number of different contents, and γ the skewness parameter. Related parameters and their typical values are depicted in Table 5.1.

Therefore, the probabilistic distributed caching problem can be simplified to one question: given the limited storage capacity and offload bandwidth, how to cache the contents across all SBSs in order to minimize the average content delivery latency?

Actually, the caching probability of the ith content can be represented as $P_i \in [0, 1]$, which means that P_i proportion of all SBSs within the cluster has cached the ith content. Then, for a ith cached content request, the local hit probability will be P_i, and the cluster hit probability P_{ic} would be $1 - P_i$ and the global hit probability will be 0. Otherwise, if the content is cached nowhere within the cluster ($P_i = 0$, $P_{ic} = 0$), then there should be a global hit from the gateway, and its probability $P_{ig} = 1$.

Besides, the corresponding latency for three different kinds of cache hit could be characterized as follows. The local hit experiences a single queue during the request, and the average processing time is $1/(\mu - \lambda_a)$, where λ_a is the sum of the original user requests rate λ and the offloaded requests rate λ_o. Specifically, the offloaded rate consists of all cluster hit requests and are evenly orchestrated to all the SBSs within the cluster, as derived in the following equation:

$$\lambda_a = \lambda + \lambda \sum_{i=1}^{N_c} P_{ic} f_i. \tag{5.2}$$

Furthermore, as stated in the delivery process, the cluster hit experiences two equivalent queues during the request, where one is in the connected SBS and the other one is in the offloaded SBS. Therefore, the average latency for the cluster hit will be double of the local hit, with quantity $2/(\mu - \lambda_a)$. Above all, the latency for the global hit is assumed to be constant.

Since each content request experiences a specific latency according to its caching strategy (i.e., the P_i), we can derive the overall average latency by multiplying the requested probability of each content by its correspondent average latency. Furthermore, the average delivery latency of a specific content is related to its hit pattern, together with the corresponding latency, as detailed in the next paragraph. The objective here is to minimize the average content delivery latency of a random user, which can be derived from the average latency of all contents in the probabilistic caching scenario. Therefore, the probabilistic caching strategy is reduced to a latency minimization problem with the variables P_i and the limited storage and bandwidth constraints as follows:

Minimize:

$$\sum_{i=1}^{N_c} f_i(P_i D_{lh} + P_{ic} D_{ch} + P_{ig} D_{gh}) \tag{5.3}$$

Subject to:

$$\sum_{i=1}^{N_c} P_i \leq C_s, \tag{5.4}$$

$$\lambda Z N_s \sum_{i=1}^{N_c} P_{ic} f_i \leq B\binom{N_s}{2}, \tag{5.5}$$

where (5.3) represents the average latency in the contents' point of view, and f_i is the requested probability of the ith most popular content.Constraint (5.4) implies the storage limitation of SBSs and constraint (5.5) means that the amount of all transferred contents within the cluster (left part) cannot exceed the overall bandwidth between all SBSs (right part) which is a combinatorial number. Here, we assume that the transferred traffic is evenly offloaded on connections between each pair of SBSs.

5.2.3 Probabilistic Cooperative Caching Strategy

Traditionally, SBSs are preferred to cache popular contents in order to reduce the overall latency. However, in addition to the storage and bandwidth constraints, the queuing delay introduced in our model makes the problem more practical but even more complicated. To solve it, we introduce a intermediate parameter S, which is the number of contents that are distributively cached in the cluster of SBSs. Hence, there are $N_c - S$ contents that should be delivered by global hit from the gateway which experiences a higher latency. Combining this definition and

the latency description given in previous sections, we can reformulate the problem into:

Minimize:

$$\sum_{i=1}^{S} f_i \frac{P_i + 2P_{ic}}{\mu - \lambda(1 + \sum_{i=1}^{S} f_i P_{ic})} + \sum_{i=S+1}^{N_c} f_i D_{gh} \tag{5.6}$$

Subject to:

$$\sum_{i=1}^{S} P_i \leq C_s, \tag{5.7}$$

$$0 \leq P_i \leq 1, \tag{5.8}$$

$$\sum_{i=1}^{S} (1 - P_i) f_i \leq B_0 = B(N_s - 1)/(2\lambda Z). \tag{5.9}$$

Given S, it can be shown that the average latency in (5.6) is increasing with the left part of (5.9), which is the portion of requests that needs to be offloaded. Therefore, to achieve the minimal latency, there are two steps to proceed: first is to maximize the intermediate variable S while fulfilling all these constraints, since the more contents cached in the SBSs, the smaller the average latency to fetch a random content. Secondly, for each S, proper values (integral multiple of $1/N_s$) are assigned to P_i of all S contents in order to reach the minimum of (5.6).

Actually, since the sequence $\{f_i\}$ is decreasing with i, the minimal value of the left side of (5.9) (denoted as Q) is 0 when $S \leq C_s$, but it increases with S when $S \geq C_s$. Therefore, to make S maximal, we increase S gradually to assure that the minimal value of Q cannot exceed B_0. As S increases to $S+1$, a portion of $1/N_s$ should be shifted from another P_i to P_{S+1}, to make sure that $1/N_s \leq P_i$ stands for each i smaller than S. After Algorithm 1 is processed, the caching proportion of each content is derived, based on which we can calculate the average content delivery latency with (5.6).

5.2.4 Performance Evaluation

In this section, we present the numerical results of our proposed algorithm compared to the uniform caching algorithm (contents are equally cached under bandwidth constraint) and the greedy algorithm (always cache the most popular contents). In detail, we depict the average latency performance according to the typical parameter setting as presented in Table 5.1.

As seen in Fig. 5.2, the average latency increases with the request rate for all three strategies. On one hand, it's caused by the increasing of queuing delay in each SBS, while on the other

Algorithm 1 Probabilistic Strategy for Cluster Caching.

Input: $\{f_i\}$, C_s, B_0, N_s, N_c;
Output: $\{P_i\}$, S;
Initialize: $\{P_i\} = 0$, $Q = 0$, $j = 1$;
while $(j < N_s)$ **do**
  $P_j = 1$;
  $j = j + 1$;
end while
while $(Q \leq B_0 \bigwedge j < N_c)$ **do**
  $P_j = \frac{1}{N_s}$;
  $P_{C_s - \lfloor \frac{j-1}{N_s - 1} \rfloor + 1} - = \frac{1}{N_s}$;
  $Q = Q + \frac{1}{N_s}(f_{C_s - \lfloor \frac{j-1}{N_s - 1} \rfloor + 1} - f_s) + f_s$;
  $j = j + 1$;
end while
$S = j - 1$;
return $\{P_i\}$, S;

hand, it's due to the relative decrease of available bandwidth. Besides, the latency increasing of greedy strategy with respect to request rate is slower than that of uniform and proposed strategy, because of the increasing effect of offloaded traffic on the queuing delay. When the request rate is relatively low, the uniform strategy outperforms the proposed strategy since more contents are cached within the cluster, while does not increase queuing delay much.

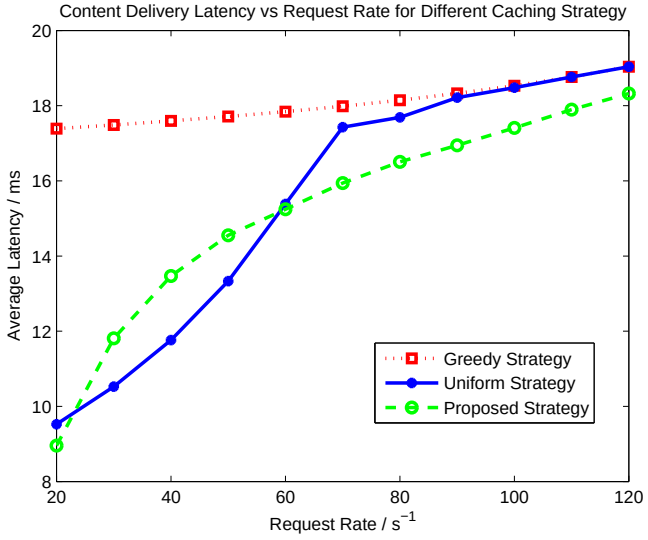


FIGURE 5.2: Average latency with respect to different request rates of each SBS.

Besides, we investigate the effect of bandwidth between SBSs on the average latency, as depicted in Fig. 5.3. Clearly, the average delivery latency of our proposed algorithm is significantly decreasing with the bandwidth, as more and more contents can be shared between different SBSs, shorting the delivery path of user requests. However, the greedy algorithm doesn't make use of the bandwidth advantage thus shows a flat line.

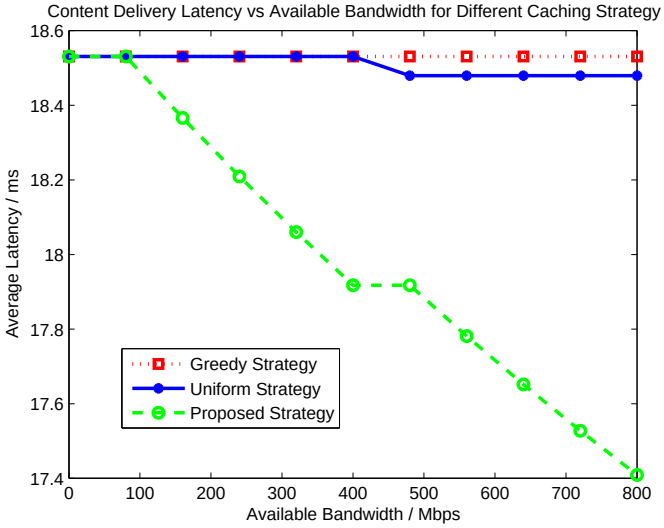


FIGURE 5.3: Average latency with different bandwidths between each SBS pairs.

Besides, all three strategies benefit from the increasing capacity of SBS, while our proposed policy outperforms the greedy one and the uniform solution as depicted in Fig. 5.4. More SBSs means more content requests can be fulfilled with cooperation thus reducing the average latency as shown in Fig. 5.5. The uniform strategy is approximated with the greedy one, which means that the available bandwidth is inadequate to store more contents than greedy strategy.

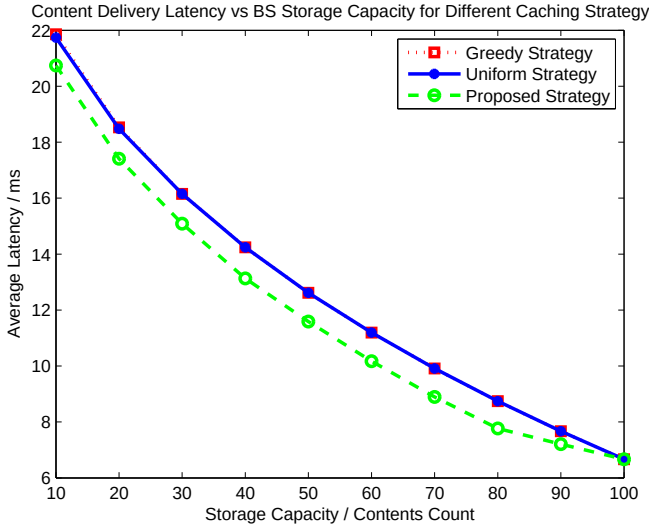


FIGURE 5.4: Average latency with different storage capacity of SBS.

After all, we investigate the effect of Zipf's skewness on the average latency performance, as shown in Fig. 5.6. Clearly, it shows that the average latencies decrease as the skewness increases for all three polices, which implies that the contents popularity are more unevenly distributed. Comparatively, our proposed algorithm always outperforms the other two with a more than 5% less latency.

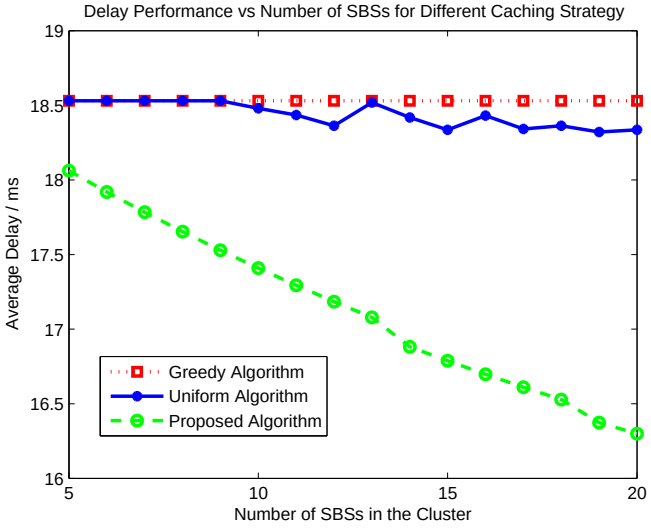


FIGURE 5.5: Average latency with different number of SBS within one cluster.

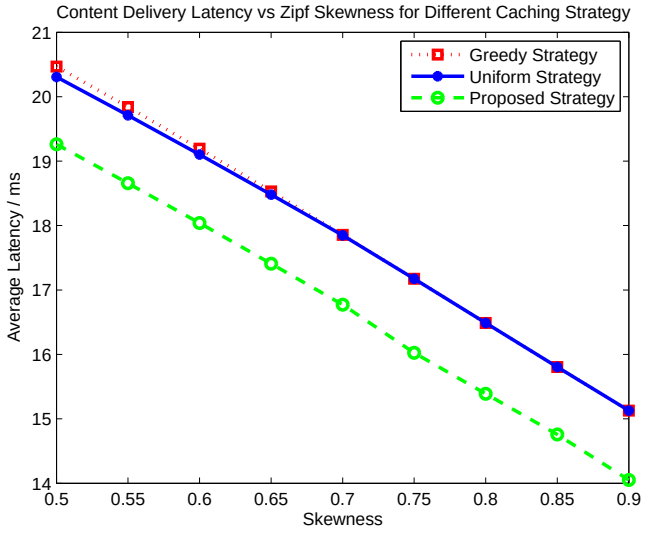


FIGURE 5.6: Average latency with different skewness of the Zipf's distribution.

In addition, as summarized in the fourth chapter of this paper through measured data, the BS's spatial density does not conform to the traditional regular hexagonal grid distribution, and it also differs greatly from the Poisson point process distribution which is often assumed in academic research. In fact, the density of BSs in urban areas tends to obey a heavy-tailed distribution, such as the α-Stable distribution. Such a spatially non-uniform BS arrangement provides favorable conditions for the mutual cooperation of neighboring BSs. Where the BS density is high, it indicates that the user density is also large, and mutual cooperation can also cache more popular content. In this way, the cooperative caching strategy proposed in this section is particularly effective in the scenario of aggregately distributed BS. Next, we analyze the influence of different parameters in the α-Stable distribution on the average latency of the cooperative caching strategy by using the BS density related parameter configuration obtained

in the previous chapter.

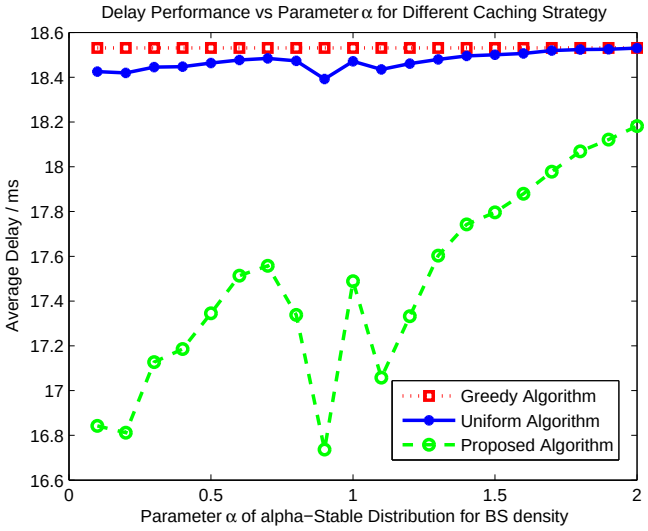


FIGURE 5.7: Average latency with different shape parameter α of the α-Stable distributed BS density.

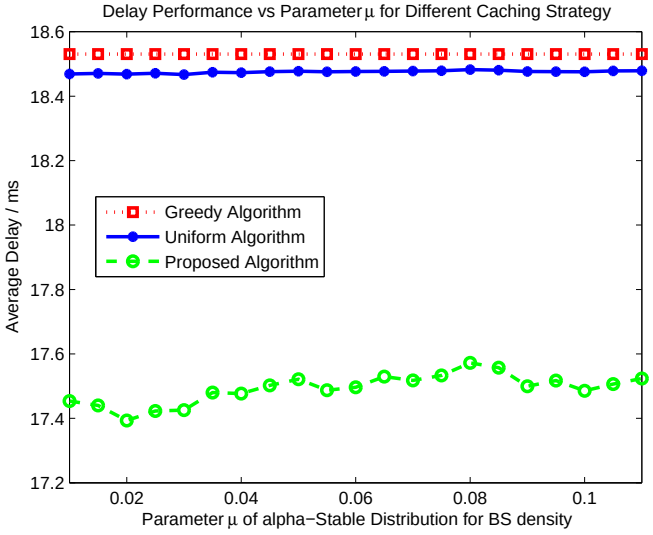


FIGURE 5.8: Average latency with different location parameter μ of the α-Stable distributed BS density.

As shown in Fig. 5.7, the latency curve of the cooperative caching strategy obviously falls below the other two curves, which shows that the proposed strategy is still valid under the heterogeneous BS distribution scenario. Specifically, as the value of α increases, the average latency of the cooperation strategy also gradually increases because the increase in the value of α generally decreases when other parameters remain unchanged. In Fig. 5.5 we have seen that the more the number of SBS in the cluster, the smaller the average latency, which explains the average latency in Fig. 5.7 increases as α value increases. As shown in the Fig. 5.8, the average latency of the cooperative caching strategy does not show a significant numerical change as the

value of μ increases, because in the case of $\alpha < 1$, the change in the μ value does not significantly change the mean value of the α-Stable distribution.

5.2.5 Conclusion

In this section, we examined the problem of probabilistic caching in a spatially aggregated cellular network scenario. Unlike the independent SBS caching in previous strategies, the transmission bandwidth between SBSs in the proposed strategy this section can be used for exchanging different contents. In order to reduce the average content delivery latency including queue waiting time, we proposed a probabilistic solution that is more efficient than the traditional greedy and uniform caching strategy. In our proposed strategy, contents with different popularity are distributed in the SBS cluster with different cache probabilities, and these cache probabilities are derived by minimizing the average latency. The numerical results of simulation tests show that, no matter in the uniform or non-uniform distribution of BSs, our proposed cooperative caching strategy can achieve smaller average content delivery latency than traditional greedy and uniform strategies under different parameters.

5.3 Clustering-oriented Multicast in Cellular Networks

The reason why multicast strategies can bring performance gains in wireless networks is due to the unique broadcasting property of wireless signals. That is, a single transmission can achieve multiple reception. Moreover, the omnipresent wireless coverage of cellular networks makes the benefit even more promising because the broadcasting gains multiplies as the coverage area and the number of users increases [43]. In practice, the multicast source (generally the BS) groups the corresponding users into the same multicast group according to the arrival time of the content request. When the number of users in the multicast group exceeds a threshold, the BS starts the multicast transmission of this content. In this way, for different request arrival patterns, the access latency experienced by each user in the same multicast group is different. In this section, based on the bursty nature of content requests derived in Chapter 4, we intend to identify how the clustering can effect traditional multicast on user's content access latency and BS's average power consumption. According to the corresponding theoretical or simulation results, a mixed unicast/multicast service mechanism with dynamic threshold is proposed.

5.3.1 Motivation and Objectives

With the exponential growth of cellular traffic, mobile users have an increasing tendency to request the same video content and the previous one-to-one transmission struggles to meet

explosive traffic demands. This provides the multicast mechanism an excellent service scenario in cellular networks. In related works, researchers focus on the problem of resource allocation between multicast and unicast transmission. In addition to the throughput which is the most intuitive indicator of the system performance, the user-perceived average latency is also a key metric as supplementary evaluation. In particular, in the performance specification for next-generation cellular networks (5G), low-latency service requirements are already one of the three major scenarios, which highlights the importance of latency-oriented architectural planning and protocol design in cellular networks. In fact, it is challenging to accurately quantify the average latency of content access in multicast scenarios. First of all, for different contents, the time taken for the BSs to acquire them from the Internet varies, while the air interface transmission time adds up to increase the diversity of the average latency. In addition, because each user's access content is different, the access frequency to the same content is also different which makes the average latency analysis of each user even more complicated. Meanwhile, due to the varieties of wireless signal environment, each user in a multicast group may experience distinctive channel environments which results in different air interface transmission rates and even the number of retransmissions. Therefore, the user's average latency analysis of multicast scenarios in cellular networks is both essential and challenging which calls for simplified assumptions about the user's request pattern, content popularity, and cable transmission delays.

In this section, we abstract the unicast/multicast process in cellular networks into a queuing model, in which all users' requests for a specific content reach a multicast group of the content. When the number of users in the multicast group is less than a prescribed threshold, the BS performs a conventional unicast process, i.e., only one user is served at a time. As new users continue to join, the BS starts a one-to-many multicast process when the number of users in the multicast group reaches the threshold. In this paradigm, different users have their own orders in the multicast group due to different arrival time which makes their probabilities of unicast-served or multicast-served also varies. In general, the average latency of mobile user and the average power consumption of the BS are two performance metrics that need to be comprehensively considered in this unicast/multicast service strategy. Here, in order to simplify the problem, we assume that each user requests for the same content library, and the popularity of different contents also meet the same Zipf distribution. In this way, given a specific multicast threshold, the serving rate of the BS and the content request arrival rate, we can utilize queuing theory to calculate the average latency of each random user from content requesting to service satisfaction.

Based on the above framework, we can first analyze the average latency for the Poisson arrival case. Different from the M/M/1[†] model in the traditional queuing theory [77], the multicast threshold parameter in our model can switch the serving rates of BSs. That is, when the queuing length is less than the threshold, BS meets the foremost user request with unicast

[†]Memoryless arrival time/Memoryless serving time/1 server.

serving rate. Once the queue length reaches the threshold, the BS will serve all the users in the queue once at the multicast serving rate. For such a unicast/multicast hybrid service paradigm under Poisson arrivals, we theoretically derived the average latency of mobile user and the multicast probability of all users, and conducted a numerical simulation accordingly for verification. Further, considering that the user's content request pattern in practice is not temporal Poisson process, the time interval of user requests is not exponentially distributed but a heavy-tailed distribution (see Chapter 4). Therefore, we need to consider the impact of the aggregation effect of user requests on the average latency of content acquisitions. In this section, we select several heavy-tailed distributions as the inter-arrival time models (such as the lognormal distribution) of user content requests. These bursty situation is compared with the ideal Poisson case through simulations to reveal the impact of aggregated nature of the content request on average latency of users and the average power consumption of BS under the unicast/multicast mixed serving mechanism.

5.3.2 Unicast/Multicast Strategy Analysis under Poisson Arrivals

In this section, we want to give a performance analysis of the mixed paradigm under different user request arrival rates through mathematical modeling. First, we characterize the arrival and departure process of content requests in the BS's unicast/multicast paradigm queue into a temporal Markov chain. Since the BS performs a multicast when the queue length reaches the threshold, the state transfer process is a circular Markov chain, as shown in Fig. 5.9.

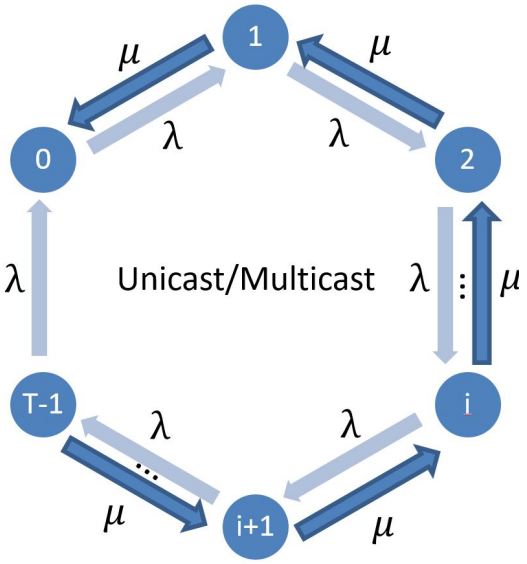


FIGURE 5.9: Markov chain characterization of the service process of unicast/multicast paradigm.

This diagram depicts the transition process of a multicast group (with threshold T) which can be thought as a close-loop Markov chain, as requests for content keep on arriving and the BS

keeps on performing unicast or multicast. In detail, the multicast queue with T possible states starts from state 0. When the first user request arrives, the Markov chain goes to state 1. In general, when the multicast queue is in state i ($0 < i < T$), it has two different ways to transfer, one of which to state $i + 1$ due to the arrival of a new content request (when $i = T - 1$, go directly to state 0), the other is transferred to the state $i - 1$ if the BS completes one unicast. The probability of each occurrence in different states depends on the content request arrival interval and the average service time of BS which is determined by the arrival rate λ and the service rate μ, respectively.

Furthermore, we hope to calculate the average latency required for a random user between sending a content request and completing the content transmission. Specifically, this problem can be divided into two separate parts. First, when a user content request arrives at a BS, it will be categorized into the service queue of the specific content. Therefore, content requests arriving at different time will encounter the service queue in different states whose probability is also the stable time proportion within the Markov chain (assuming $P_i, 0 \leq i < T$). Second, for a random request arrival, assuming that the service queue is in state i when it joins the multicast group, thus this user is listed in order i of unicast service. Because this request may be served by unicast (with unicast order i), or by multicast (triggered when queue length reaches T), we assume the average waiting time is D_i. Thus, for a random content request, its average latency is:

$$D = \sum_{i=0}^{T-2} P_i D_{i+1}, \tag{5.10}$$

where P_i is the stable probability of the Markov chain in state i and D_{i+1} the average waiting time of the user who arrives the queue at state i. The reason why the index is $i+1$ here is that the Markov chain status will be shifted from i to $i + 1$ after the new user joins. Next, we need to calculate P_i and D_i separately.

5.3.2.1 Average Request Latency under Poisson Arrivals

Here, in order to simplify the theoretical derivation of the average latency, we assume that the arrivals of content requests obeys the Poisson process. First, we derive the stable distribution of the Markov chain in Fig. 5.9 based on the content request's arrival rate λ, the BS unicast service rate μ, and the multicast threshold T (as Markov chain with finite states should have limiting probabilities). According to the rate principle, we can obtain the following equilibrium equations:

$$\begin{aligned} P_1(\lambda+\mu) &= P_0\lambda + P_2\mu, \\ P_i(\lambda+\mu) &= P_{i-1}\lambda + P_{i+1}\mu, \\ P_{T-1}(\lambda+\mu) &= P_{T-2}\lambda, \\ P_0\lambda &= P_1\mu + P_{T-1}\lambda. \end{aligned} \tag{5.11}$$

For example, the left side of the first equation represents the departure rate of state 1 and the right two parts are the arriving rates from state 0 and state 2 to state 1 respectively. The same equation also applies to state i, when $0 < i < T-1$. In particular, if the queue is in state $T-1$, then the departure rate is $P_{T-1}(\lambda+\mu)$, and the arrival rate is $P_{T-2}\lambda$. Similarly for state 0, the departure rate is $P_0\lambda$ and the arrival rate is from state 1 for unicast and from state $T-1$ for multicast.

Combine the above equilibrium equations with the definition that limiting probabilities sum up to 1:

$$\sum_{i=0}^{T-1} P_i = 1. \tag{5.12}$$

When $\lambda \neq \mu$, we can get the following non-trivial solution:

$$P_i = c_1(\frac{\lambda}{\mu})^i + c_2, \tag{5.13}$$

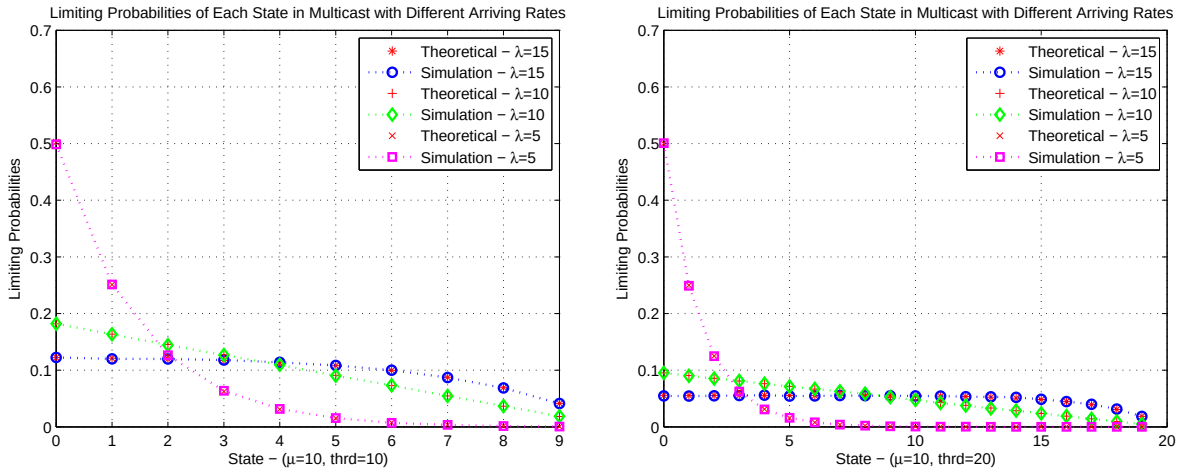
where the values of c_1 and c_2 are calculated as follows.

$$c_1 = \frac{1}{\frac{1-(\frac{\lambda}{\mu})^T}{1-\frac{\lambda}{\mu}} - T(\frac{\lambda}{\mu})^T}, c_2 = -c_1(\frac{\lambda}{\mu})^T. \tag{5.14}$$

When $\lambda = \mu$, we can get the following trivial solution:

$$P_i = \frac{2(T-i+1)}{T(T+1)}. \tag{5.15}$$

Therefore, we theoretically present the limiting probabilities of the single-content service queue under the unicast/multicast hybrid service mechanism. To verify the correctness, we will next calculate the corresponding theoretical and simulation values of the Markov chain with different parameter settings (λ, μ, T).



(a) Limiting probabilities of each state in the Markov chain when threshold $T = 10$.

(b) Limiting probabilities of each state in the Markov chain when threshold $T = 20$.

FIGURE 5.10: Limiting probabilities of each state in the Markov chain under different Poisson arrival rates.

From Fig. 5.10 we can see that for $\lambda < \mu$ ($\lambda = 5$, $\mu = 10$), the limiting probabilities of the Markov chain decreases as value i grows. Since the unicast service rate of the BS is obviously greater than the arrival rate of user content request, most of the content requests are served in time through the unicasts, thus the service queue rarely enters the multicast mode. For the case of $\lambda > \mu$ ($\lambda = 15$, $\mu = 10$), the limiting probabilities of the Markov chain also follows the same decreasing pattern, while the rate of decrement is increasing, which is opposite with the $\lambda < \mu$ case. In addition, for $\lambda = \mu$, the limit probabilities show a linear decreasing trend. Besides, the simulation curves in those figures perfectly match with the corresponding theoretical curves which certifies the correctness of our derivation. Therefore, we have finished the derivation of limiting probability in the service queue, and then we need to complete the average waiting time part.

Since BS's serving procedure is assumed here to be a mixed process of unicast and multicast, the average waiting time for a user in the service queue is not only related to the queue length which affects the multicast time, but also related to the unicast order which determines the unicast time. Thus, we can use a tuple $(m, \quad n)$ to fully describe any user in the service queue, where m represents the unicast order, and n means that the queue currently requires n requests to trigger the multicast, indicating that there are $T - n$ users waiting in line. Meanwhile, based on the definition of m and n above, we can draw the constraint $m \leq T - n$, equally $m + n \leq T$. Furthermore, we hope to derive the recursion pattern and then the general formula of the user's average wait time $W(m, n)$ with parameters (m,n,T) according to the process in Fig. 5.9.

Fig. 5.9 describes the transition process of the queue length in the Markov chain. In most cases, each transition in the chain may refer to queuing length plus one due to the arrival of

new request or queuing length minus one due to unicast of the foremost request. Combined with the $(m,\ n)$ description, the user's unicast order remains (m unchanged) when a new request arrives, while the number of new requests necessary for the queue to trigger multicast become $n-1$, then the tuple becomes $(m,\ \ n-1)$. When a unicast is completed, the unicast order of the user will become $m-1$ ($m>1$, otherwise it will be served immediately), and the number of requests necessary for multicast will be $n+1$, so the state tuple becomes $(m-1,\ n+1)$. According to the above description, we can obtain a two-dimensional state transition diagram as shown in Fig. 5.11.

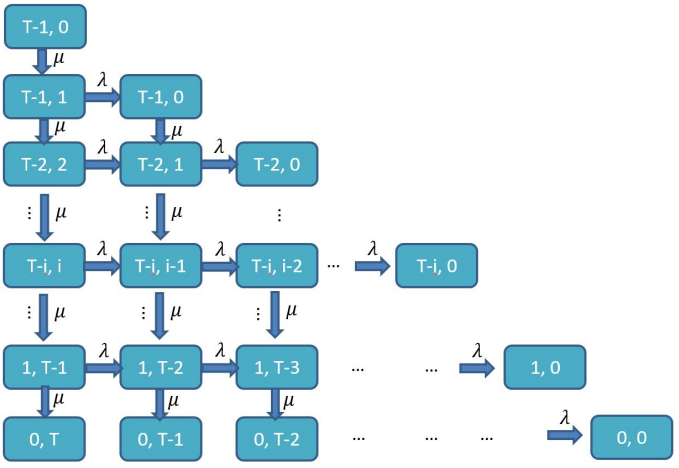


FIGURE 5.11: Two-dimensional state transition diagram illustrating unicast/multicast service queuing.

Combining the above description of the state transition with Fig. 5.11, we can obtain the following recursion pattern for $W(m,n)$:

$$W(m,n)=\frac{1}{\lambda+\mu}+\frac{\lambda}{\lambda+\mu}W(m,n-1)+\frac{\mu}{\lambda+\mu}W(m-1,n+1), \tag{5.16}$$

where m and n meet the constrains $m+n\leq T$, $m>0$, and $n>0$. Besides, the boundary values for $W(m,n)$ satisfy

$$W(m,0)=\frac{1}{\mu},\qquad W(0,n)=0. \tag{5.17}$$

Based on the above recursion formula and boundary conditions, we can derive the general formula of $W(m,n)$:

$$W(m,n)=\frac{1}{\mu}-\frac{\mu}{(\lambda+\mu)^2}+\frac{1}{\lambda+\mu}\sum_{i=0}^{m-1}\sum_{j=0}^{n+i-1}(\frac{\mu}{\lambda+\mu})^i(\frac{\lambda}{\lambda+\mu})^j\binom{i+j}{i}, \tag{5.18}$$

where $\binom{i+j}{i}$ is the combinatorial number with value $\frac{(i+j)!}{i!j!}$. In this way, we formulate the average waiting time for an arbitrary user with tuple state $(m, \ n)$ in a multicast queue. In order to calculate the average waiting time D_{i+1} for the new request in Eq. 5.10, we need to examine the relationship between D_i and $W(m,n)$. According to the previous definition, D_i represents the average waiting time of users who entered the queue in the unicast order of i, while $W(m,n)$ represents the average waiting time for users in the m unicast order with queue length $T-n$. Obviously, according to their respective definitions, we can get the following equation:

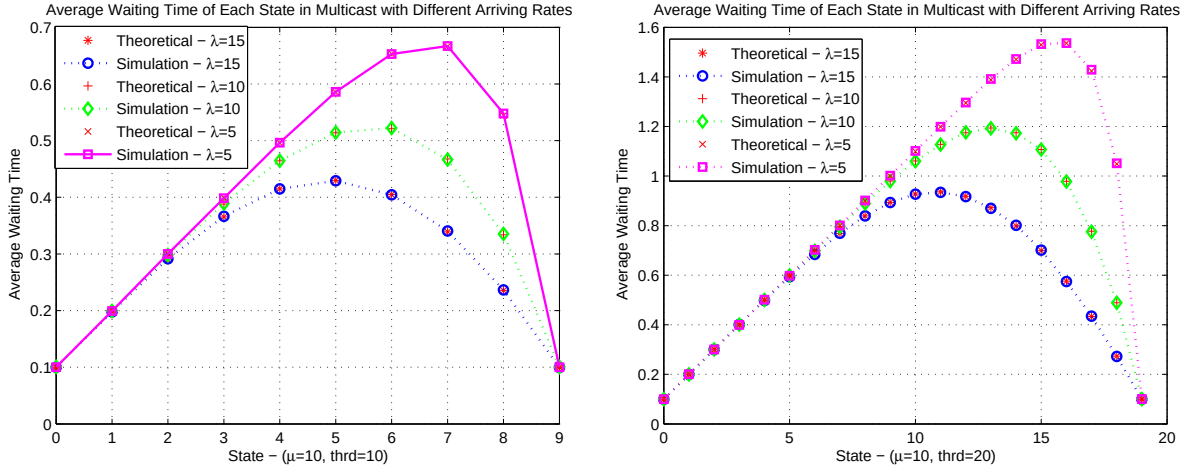
$$D_m = W(m, T-m). \tag{5.19}$$

Combining with the general term in Eq. 5.18, we can get the mathematical expression of D_m:

$$D_m = \frac{1}{\mu} - \frac{\mu}{(\lambda+\mu)^2} + \frac{1}{\lambda+\mu}\sum_{i=0}^{m-1}\sum_{j=0}^{T-m+i-1}(\frac{\mu}{\lambda+\mu})^i(\frac{\lambda}{\lambda+\mu})^j\binom{i+j}{i}. \tag{5.20}$$

After combining Eq. 5.13 of P_i with Eq. 5.20 of D_i according to Eq. 5.10, we can derive the average latency of a random user with a unicast/multicast hybrid strategy under Poisson's assumption. In order to verify the correctness of the above theoretical derivation, we also conduct simulation under different parameter configurations.

First of all, for the average latency of different orders in the service queue, we derive D_i from Eq. 5.20 for several combinations of arrival and service rates, and the theoretical value and corresponding simulation curves are shown in Fig. 5.12.



(a) Average waiting time for different states when threshold $T = 10$.

(b) Average waiting time for different states when threshold $T = 20$.

FIGURE 5.12: Average waiting time for each state in the Markov chain under different parameter setups.

From Fig. 5.12, we can see that for different parameter configurations, D_i first increases and then decreases as state order i grows. Specifically, for the case of $\lambda \leq \mu$, the increasing rate of D_i when i is small is significantly less than that when i is greater; for case $\lambda > \mu$ ($\lambda = 15$, $\mu = 10$), the increasing rate of D_i when i is smaller is comparable to the decreasing rate when i is larger. This may be due to the fact that when $\lambda \leq \mu$, more content requests are unicast served, so the maximum value of D_i should be obtained within the range of $i > T/2$, thus D_i curve shows a right shoulder shape. For $\lambda > \mu$, content requests are relatively less unicast served, so the peak of the curve is shifted left. In addition, for different parameter configurations, the theoretical and simulation curve are well matched, which also explains the correctness of the previous formula derivations.

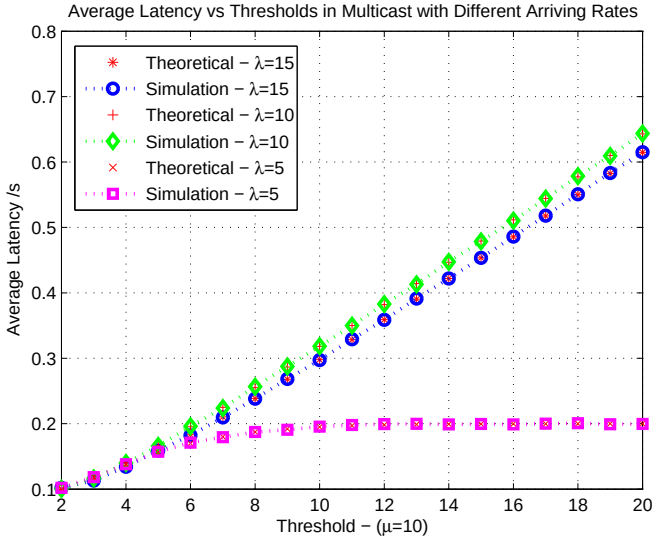


FIGURE 5.13: Average latency varies with multicast threshold T under different arrival rates.

Similarly, for the average waiting time D of random user under Poisson arrivals, we plotted Eq. 5.10 and the simulated D for different combinations of arrival and service rates. The value varies with the multicast threshold T, as shown in Fig. 5.13 while service rate is fixed at $\mu = 10$. For arrival rates of $\lambda = 15$, $\lambda = 10$, and $\lambda = 5$, we give the theoretical and simulated average latency with the multicast threshold respectively. As seen from Fig. 5.13, the average latencies increase with the threshold value in all cases. When $\lambda \geq \mu$, the average latency shows a linear upward trend, and the average latency curve under $\lambda = 15$ falls below the curve of $\lambda = 10$. Since the multicast probability increases as $\lambda > \mu$, the content request does not need to wait for an one-to-one unicast transmission; when $\lambda < \mu$, the average latency gradually increases if $T < 10$, and remains constant after $T > 10$. This result shows that regardless of the relationship between the arrival rate of the content request and the service rate of a single BS, the average latency of request in the unicast/multicast policy always increases with the multicast threshold.

5.3.2.2 Average Power Consumption of BS under Poisson Arrivals

In the study of the unicast/multicast hybrid strategy, the power consumption of BS is equally important as the average latency of requests. The multicast process essentially utilizes the temporal aggregation characteristics of user requests to exchange spectrum and energy efficiency with latency costs since one-time multicast of BS can satisfy the content requests of T users. Therefore, we hope to further examine the average power consumption of BS in the hybrid service strategy by firstly analyzing the overall multicast probability of the unicast/multicast service queue. As shown in Fig. 5.14, we plot the multicast probability for λ and μ in three cases. In this figure, three curves collectively move up as λ increases, i.e., the multicast probability increases. From another point of view, as T increases, the multicast probability decreases and predictably tends to be flat, because the threshold becomes harder to reach thus more content requests are unicast served.

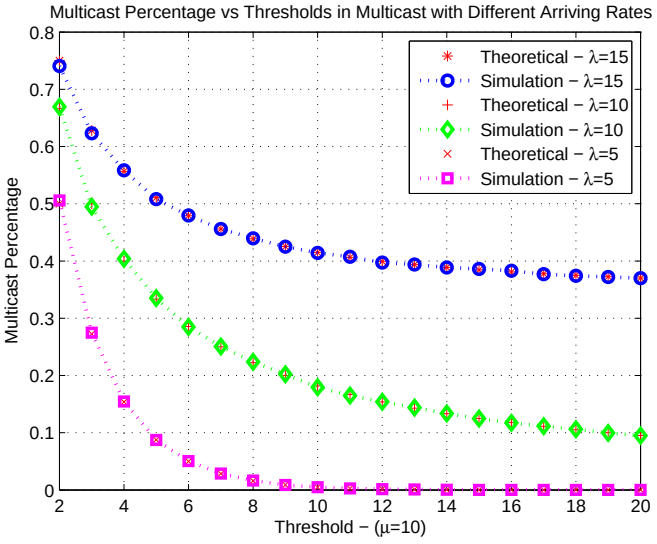


FIGURE 5.14: Multicast probability varies with threshold T under different arrival rates.

Starting from the multicast probability, we plot the average power consumption of BS as a function of T in Fig. 5.15. For the case $\lambda \leq \mu$, as multicast probability decreases, more and more users need to be served through unicast, thus the average power consumption also increases with the value of T. While for $\lambda > \mu$, the average power consumption decreases with T because the multicast probability decreases more slowly than the increasing of T. If λ far exceeds μ, the service queue is in a congested status, and the average latency of requests and the average power consumption of BS increases and decreases with T, respectively, thus we need to make a trade-off in the help of the multicast threshold.

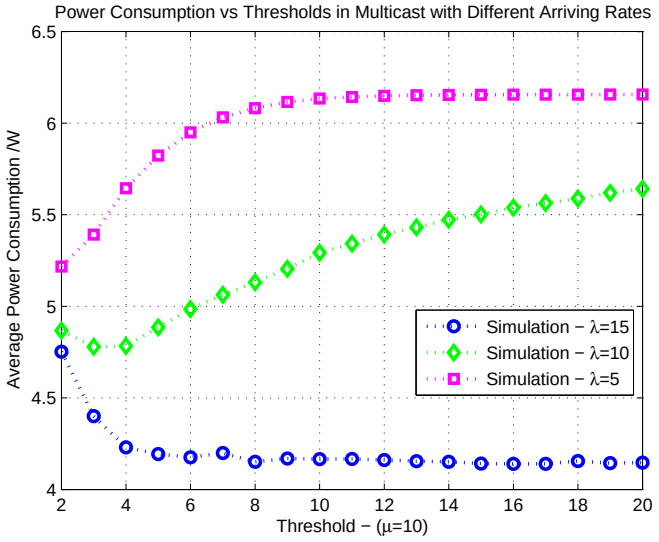


FIGURE 5.15: Average power consumption varies with multicast threshold T under different arrival rates.

5.3.2.3 Joint Optimization between Request Latency and Power Consumption

Since the average latency of request and the average power consumption of BS vary oppositely with the multicast threshold, we need to select a trade-off strategy combining these two optimizing goals. Firstly, we introduce a normalization parameter ϵ, which combines latency D and power consumption P into an overall evaluation metric $(D + \epsilon P)$. Fig. 5.16 depicts the relationship between multicast threshold T and the joint metric with $\epsilon = 1$ for different values of λ. We can see that the curve with $\lambda = 5$ is at the top and keeps growing, while the other two curves decreasing first and then increasing, and the lowest point of the $\lambda = 15$ curve is comparably right shifted. For different values of ϵ on $(0, 1)$, we represent the corresponding optimal multicast threshold in Fig. 5.17. When $\lambda = 5$, the optimal threshold is always 2; as λ increases, the curve shows a stair-like rise since the increase of ϵ strengthens the importance of average power consumption in the joint optimization, so the optimal strategy tends to reduce P by increasing T.

Through the above analyses, we studied the performance of the unicast/multicast hybrid transmission strategy under Poisson arrivals, including the average latency of request and the average power consumption of BS. As the theoretical derivation and corresponding simulation results show, the newly proposed hybrid transmission strategy can well solve the problem that the user latency in the congestion scenario $(\lambda > \mu)$ tends to be infinite, and it can also partly reduce the average power consumption of BS. In such a hybrid strategy, we found that the average latency of request does not necessarily increase with the arrival rate λ, while sometimes achieving a smaller latency with higher arrival rate due to multicast. Meanwhile, the average power consumption of BS also exhibits different characteristics depending on λ and μ. In addition,

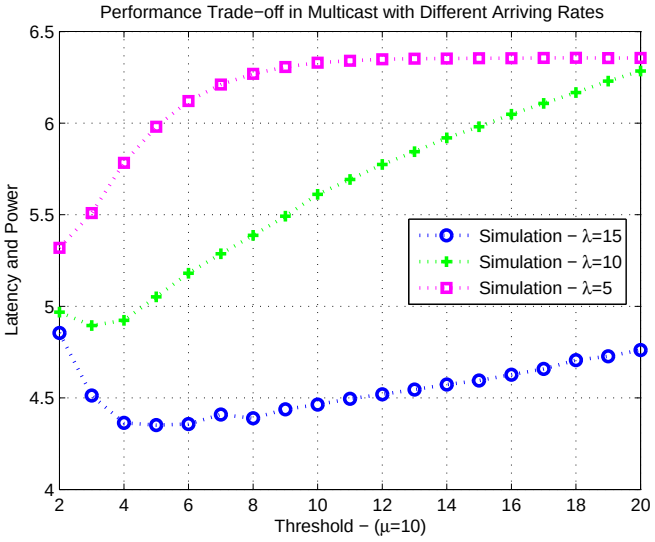


FIGURE 5.16: Latency and power trade-offs with multicast threshold T under different arrival rates.

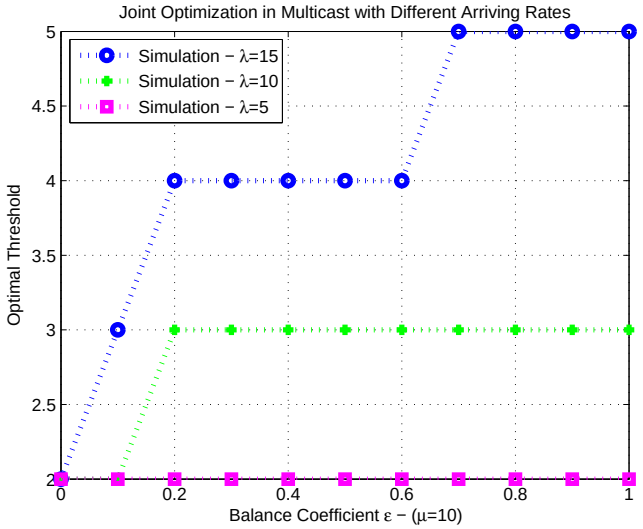


FIGURE 5.17: Optimal multicast thresholds for joint optimization of latency and power under different arrival rates.

we found that in different parameter setups, this hybrid strategy has a corresponding optimal multicast threshold to minimize the weighted sum of the average user latency and the average power consumption of BS. In this way, the BS needs to be dynamically configured according to the arrival pattern of content requests to optimize the overall performance.

5.3.3 Unicast/Multicast Strategy Analysis under Bursty Arrivals

The previous section explains how average latency/power consumption vary with relevant parameters in the unicast/multicast scenario under Poisson arrivals. However, in fact, users'

content request pattern in cellular networks is far from the traditional Poisson process. That is, the inter-arrival time between requests does not obey the exponential distribution, but may obey the lognormal distribution as obtained from real data analysis in Chapter 4. Therefore, in this section, we will consider the impact of the burstiness of content requests on the average latency and power consumption.

Different from the Poisson arrivals of content requests, the performance metric in the unicast/multicast policy cannot be theoretically derived under bursty arrivals, such as the limiting probabilities and average waiting time of different states in service queue, or the multicast probability for each request. The reason why closed-form solution is not derivable here is that the Markov chain can't be expressed as a stable transfer process. Therefore, in the following performance analysis of non-Poisson arrivals, we mainly compare different situations through numerical simulations.

5.3.3.1 Average Latency of Request under Bursty Arrivals

First of all, for different possible arrival patterns, we not only analyze the effect of multicast thresholds on the average latency, but also conduct the comparison between different request arrival rates. Besides the exponential distribution and lognormal distribution, we have chosen uniform distribution and semi-Gaussian distribution as candidate arrival patterns. The reason behind these selections not only includes their universality in practice but also their mathematical properties. Actually, in order to reflect the impact of statistical distribution, it is necessary to ensure that the mean and variance of those selected distributions are equal. The PDFs of those statistical distributions will be briefly described below. In simulation, we will make sure that the combinations of the mean and variance of those distributions are consistent.

For convenience, we use t here to indicate the inter-arrival time of random content requests in a content service queue.

- If the PDF of random variable t satisfies:

$$f(t) = \frac{1}{b-a}, \qquad a \leq t \leq b. \tag{5.21}$$

  Then t obeys the uniform distribution with parameters (a,b), and its mean is $(a+b)/2$ while variance is $(b-a)^2/12$.

- If the PDF of random variable t satisfies:

$$f(t) = \frac{\sqrt{2/\pi}}{\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \qquad t > \mu. \tag{5.22}$$

Then t obeys the semi-Gaussian distribution with parameters (μ,σ), and its mean is $\mu + \sigma\sqrt{2/\pi}$ while variance is $\sigma^2(1 - \frac{2}{\pi})$.

From Fig. 5.18, we can see that when (λ, μ)=(15, 10), the average latency for different arrival patterns exhibits a similarly linear growth, and the average latency of exponential case is obviously higher than that of others. Again in Fig. 5.19, the theoretical and simulation values of the multicast probability under exponential interval are well-matched and slightly larger than others. Among them, multicast probability of the uniform distribution is the lowest, and the semi-Gaussian distribution follows as second. Different from the previous two figures, the variation of the average power consumption of BS in Fig. 5.20 with the multicast threshold can clearly distinguish these four different distributions, where the uniform distribution occupies the largest, the semi-Gaussian distribution is second and the curves of lognormal distribution and exponential distribution twist. From these three figures, we can draw the following conclusion: even if different arrival patterns have the same mean and variance, they still have different performance in the unicast/multicast hybrid strategy. Next, we will focus on the lognormal distribution which is more consistent with the real data, and examine the impact of different parameters on average latency and power consumption.

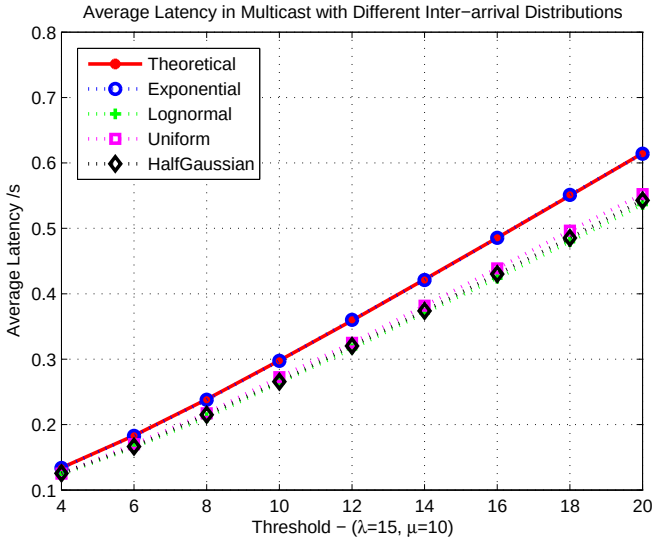


FIGURE 5.18: Average latency varies with multicast threshold T in different request inter-arrival distributions.

In exponential case, the mean and standard deviation of the random variable are equal to $1/\lambda$. Therefore, once λ is determined, the mean and variance cannot be adjusted separately for exponential distribution. However, in practice, same number of requests can show a variety of statistical distributions as the arrival pattern changes, which is consistent with the dynamics of mobile content requests. From this point of view, the exponential distribution (Poisson arrival process) does not have the flexibility to describe the temporal pattern of content requests, as we revealed in previous chapters. On the other hand, the lognormal distribution has more flexibility

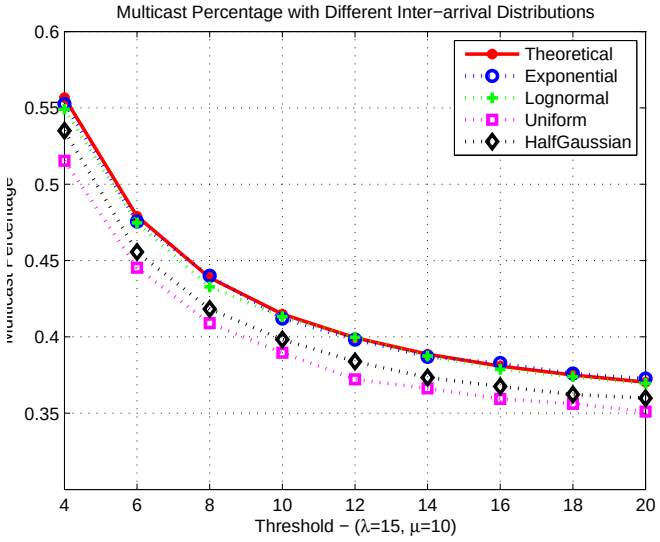


FIGURE 5.19: Multicast probability varies with multicast threshold T in different request inter-arrival distributions.

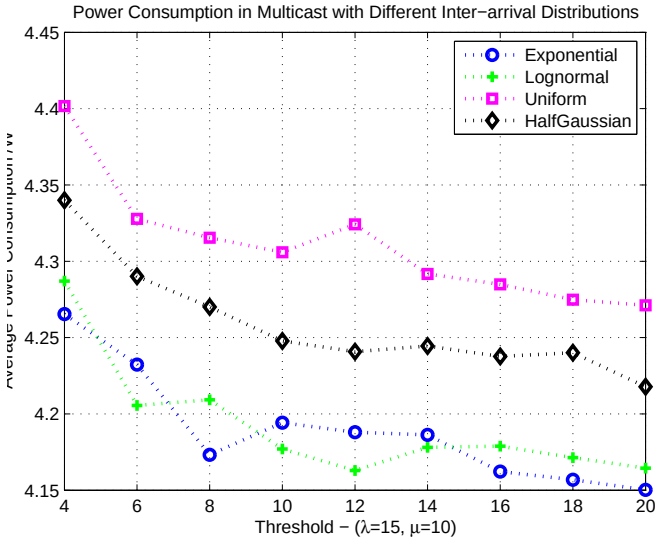


FIGURE 5.20: Average power consumption varies with multicast threshold T in different request inter-arrival distributions.

in parameter selection since it can adjust the variance while maintaining the mean value, which is useful to simulate the burst characteristics of request arrivals. Next, we analyze the numerical impact of burstiness on latency performance and energy efficiency in unicast/multicast strategies by adjusting the ratio of the standard deviation to the mean in lognormal distribution (ρ, which is used to characterize the degree of temporal aggregation of content requests, as defined by Eq. 5.23). .

$$\rho = \frac{\sqrt{Var(t)}}{E(t)}. \tag{5.23}$$

From Fig. 5.21 we can see that, the average latency of the lognormal distribution with the same degree of aggregation ($\rho = 1$) is smaller than that of exponential distribution; while among different lognormal distributions, it shows that the greater ρ, the smaller the average latency. A possible explanation is that the increase in ρ indicates greater variance in the arrival time interval for the same number of requests which results in more burstiness, thus the requests in congested state is mostly served by multicast, while the requests in idle state are served through unicast. The superposition of the two cases degrades the overall average latency. Such an explanation can also be confirmed by the curves in Fig. 5.22 where the overall multicast probability of the service queue clearly increases with ρ.

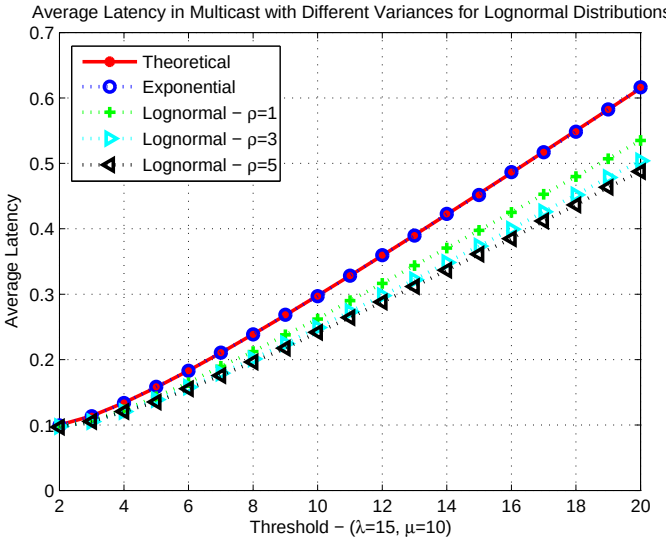


FIGURE 5.21: Average latency varies with the multicast threshold T under lognormal distribution with different degrees of aggregation ρ.

5.3.3.2 Average Power Consumption of BS under Bursty Arrivals

In order to examine the variation of the average power consumption of BS with the degree of aggregation ρ under bursty arrivals, the multicast probability of the service queue should be analyzed firstly. In Fig. 5.22, the lognormal distribution and exponential distribution with the same ρ have no significant difference in the multicast probability. However, the difference between lognormal distributions with distinct ρ is obvious, which also shows the significance of this aggregation factor in the analysis of multicast probability. Further, for Fig. 5.23, consistent with the previous discussion, as the multicast probability increases, BS can serve most users with less power loss through more multicast and the average power consumption also decreases as ρ increases. As the multicast threshold increases, no matter the requests are exponentially or lognormally arrived, the average latency always increases while the average power consumption decreases mostly. From this point of view, we can select a joint optimal solution between the two metrics by adjusting T.

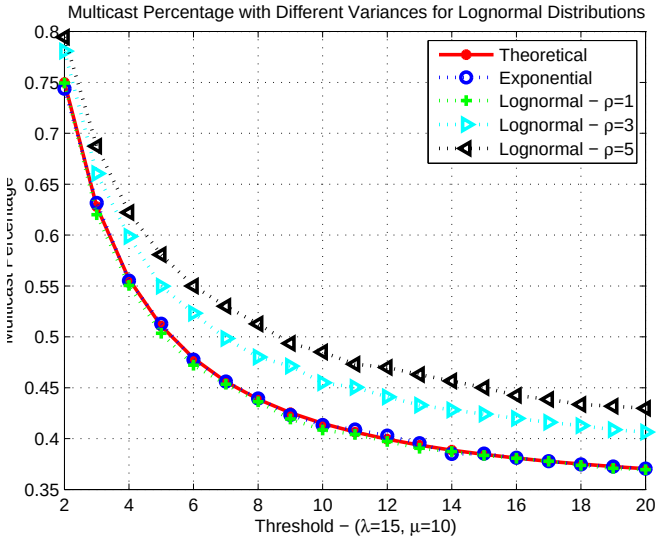


FIGURE 5.22: Multicast probability varies with the multicast threshold T under lognormal distribution with different degrees of aggregation ρ.

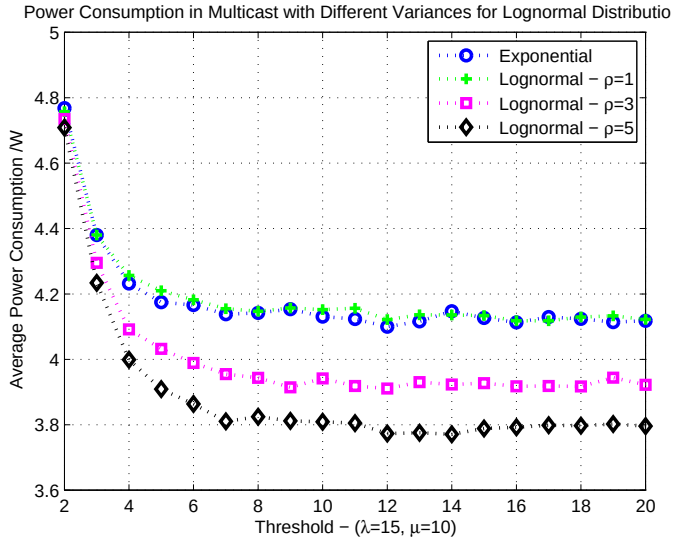


FIGURE 5.23: Average power consumption varies with the multicast threshold T under lognormal distribution with different degrees of aggregation ρ.

After analyzing the impact of aggregation degree on the performance of the unicast/multicast hybrid strategy in a lognormal distributed time interval, we return to analyze the impact of arrival rates. Specifically, we fix $\rho = 1$ so that a unified comparison can be performed on λ. Unlike the consistent increase of average latency with respect to the arrival rate λ as shown in Fig. 5.13, we can see that the average latency under lognormal distributions shows an overall downward shift as λ increases in Fig. 5.24. For example, for $\lambda = 15$, the average latency under the lognormal distribution is not only smaller than that of the exponential distribution with the same arrival rate, but is also significantly lower than the lognormal case of $\lambda = 10$. This result shows that the content requests under lognormal arrival is more suitable to be

served by multicast strategy than those under exponential distribution. In other words, the unicast/multicast hybrid strategy we proposed here is very effective to reduce the average latency under bursty arrivals which is very challengeable only by unicast.

Correspondingly, Fig. 5.25 shows that the multicast probability under lognarmal distributions increases significantly with the arrival rate, such as the curve of $\lambda = 15$ is generally higher than that of $\lambda = 10$ and $\lambda = 5$, and it is also in good agreement with the theoretical and simulation curves of corresponding exponential distribution. In addition, each multicast probability curve gradually decreases as T increases, but the declining becomes slower and slower. This is due to the fact that as T increases, the service queue needs more content requests to trigger the multicast. As T continues to increase, although the number of multicast decreases, the number of users being served in each multicast increases which contributes to the flat line in the end.

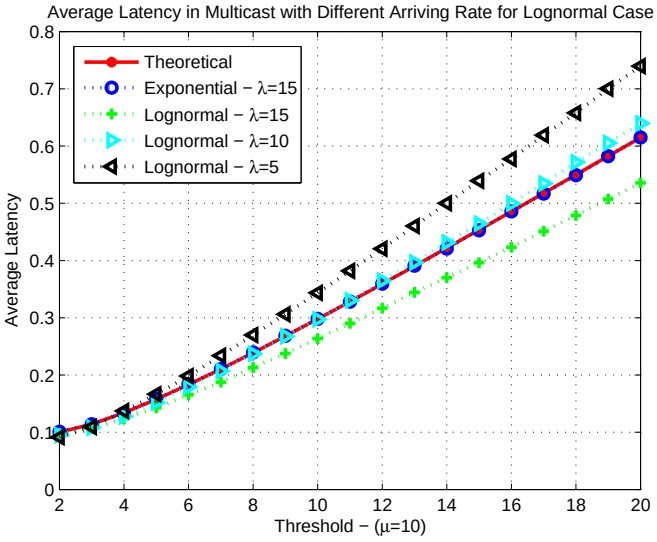


FIGURE 5.24: Average latency varies with the multicast threshold T under lognormal distribution with different average arrival rate λ.

We further investigate the variation of average power consumption with the multicast threshold at different arrival rates as depicted in Fig. 5.26 where average power curve shows three different trends. Specifically, when $\lambda = 15 > \mu$, the curve shows a gradual decrease as T increases; while in $\lambda = 10 = \mu$, the curve of lognormal distribution shows a trend of decreasing first and then increasing and an enduring increase is found in $\lambda = 5 < \mu$ case. In fact, the overall average power consumption and multicast probability of the unicast/multicast service strategy are highly related. If one request is unicast served, the amount of power it consumes (assumed to be random variable P_1) is determined by the given SIR, channel conditions and the user's distance from the BS. If this request is being served by multicast, then the BS consumes the maximum power (P_{max}) required by T users in this multicast group, and each single user only consumes $\frac{P_{max}}{T}$ of the power. Therefore, combined with the multicast probability M_T, the average power consumption of BS for one single request can be written as:

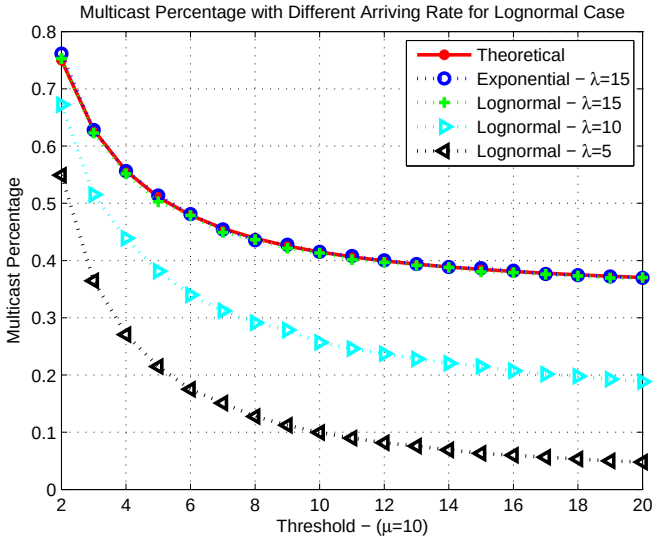


FIGURE 5.25: multicast probability varies with the multicast threshold T under lognormal distribution with different average arrival rate λ.

$$P = (1 - M_T)P_1 + M_T \frac{P_{max}}{T}. \tag{5.24}$$

Since both P_1 and P_{max} are independent of T, and their values are in the same order of magnitude, for simplicity we assume them to be the same parameter P_0, representing the BS's transmission power each time. So the Eq. 5.24 can be simplified to:

$$P = (1 - \frac{T-1}{T} M_T)P_0. \tag{5.25}$$

From Eq. 5.25 we can see that the variation of average power consumption P with respect to the threshold T is reflected in two aspects where $\frac{T-1}{T}$ increases with T but M_T decreases with T as shown in Fig. 5.25. Taken together, when the increasing of $\frac{T-1}{T}$ prevails over the decreasing of M_T, P will show a decline on T. As in Fig. 5.25, the larger λ, the slower the decreasing of M_T with T. The first half of the $\lambda = 10$ curve and the $\lambda = 15$ curve in Fig. 5.26 both show a downward trend since $\frac{T-1}{T}$ increases more rapidly when T is small. However, the opposite situation is true for the second half of the $\lambda = 10$ curve and the $\lambda = 5$ curve in Fig. 5.26.

5.3.3.3 Joint Optimization of Latency and Power Consumption under Bursty Arrivals

Similarly, after analyzing each performance metric separately, we hope to comprehensively examine the average latency of request and the average power consumption of BS, and perform the optimization of $D + \epsilon P$ on the selection of multicast threshold T. As shown in Fig. 5.27,

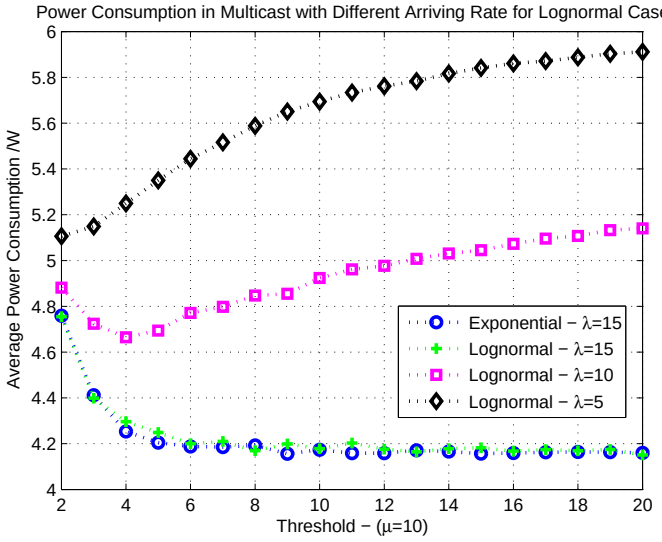


FIGURE 5.26: Average power consumption varies with the multicast threshold T under lognormal distribution with different average arrival rate λ.

we show the joint metric with T for the lognormal distributed inter-arrival time with different arrival rates given $\epsilon = 1$. The three curves all show a decreasing-then-increasing trend except that $\lambda = 5$ curve keeps rising. As a result, each curve has a lowest point minimizing the target metric. Specifically, curves continue to decrease as λ increases, and the optimal T value also gradually increases since the average power consumption decreases with T, thereby reducing the joint metric. The reason why different curves show a similarly decreasing-then-increasing trend is the same with the explanation of Fig. 5.29.

In order to examine the impact of coefficient ϵ on the joint optimization results, we depict the variation of the optimal threshold T with ϵ in (0,1) for different arrival rates in Fig. 5.28. It can be seen that regardless of the specific arrival interval, the optimal T value of the joint optimization shows a stair-like upward trend with increasing ϵ except for the $\lambda = 5$ case. For example, the optimal value in the $\lambda = 15$ case of a lognormal distribution, increases gradually from $T = 2$ in $\epsilon = 0$ to $T = 6$ when $\epsilon = 1$. The possible explanation is that, as ϵ increases, larger proportion in the joint optimization lies on the average power consumption which generally shows a decreasing-then-increasing trend with respect to T. In order to minimize the joint metric, T needs to be around the lowest point of average power consumption curve. Therefore, as the value of ϵ increases, the optimal threshold T will keep increasing at first. However, once the value of T reaches the lowest point, it remains constant since both the average latency and power consumption will increase as T increases.

Furthermore, besides the arrival rate, we analyzed the effect of the degree of aggregation ρ on the joint optimization performance. As shown in Fig. 5.29, we show the variation curve of the joint metric with multicast threshold for different ρ value given $\epsilon = 1$, which is also a decreasing-then-increasing trend. The possible explanation is that when T is small, the multicast effect is

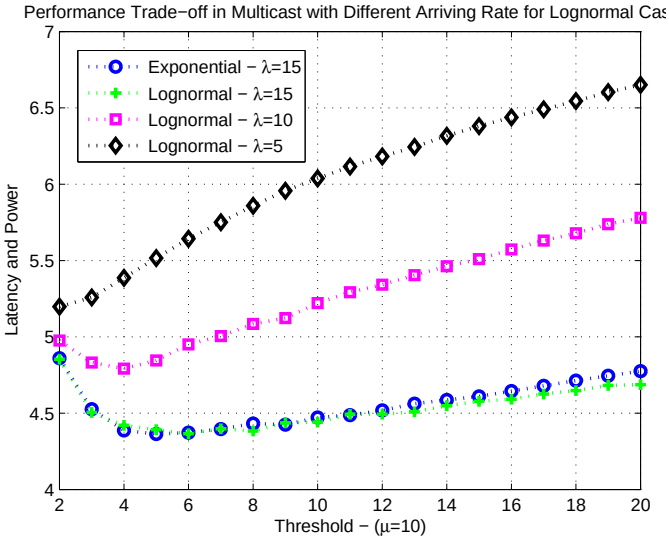


FIGURE 5.27: Latency and power trade-offs with multicast threshold T under different arrival rates for lognormal distributed inter-arrival time.

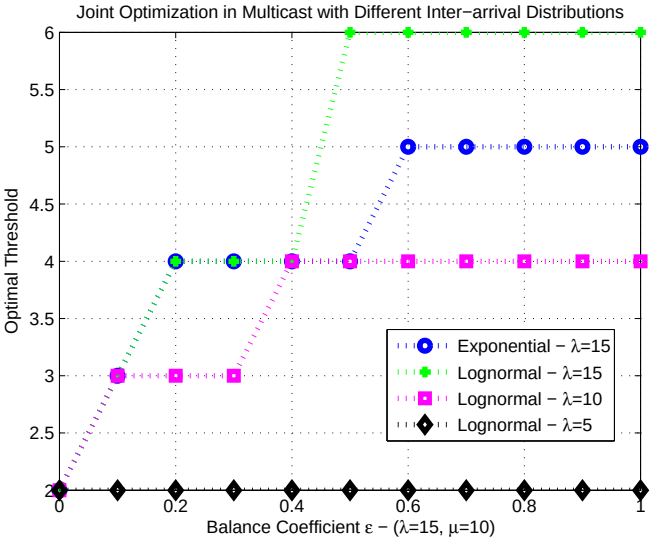


FIGURE 5.28: Optimal multicast thresholds for joint optimization of latency and power under different arrival rates for lognormal distributed inter-arrival time.

significant and the average power consumption decreases rapidly with T where the rate of decline exceeds the growth rate of average latency. When T exceeds a certain value, the decreasing of multicast probability makes the average power consumption no longer significantly drops. On the other hand, the average latency will play a dominant role in the joint metric variation, thus the segment increases with T. Through comparison, it is found that different inter-arrival time corresponds to different optimal multicast threshold even with the same arrival rate and service rate. For example, in Fig. 5.29, the optimal value of the exponential case is $T = 5$, while the optimal T values of the other three lognormal curves increase with ρ. In addition, we can obtain different optimal multicast threshold T when choosing different ϵ values to jointly optimize the

average latency and power consumption. Specifically, when ϵ is within the range of (0,1), the variation curve of optimal T value for different inter-arrival distributions is shown in Figure 5.30. It can be seen that regardless of the specific choice of the arrival pattern, the optimal T value of the joint optimization shows a stair-like upward trend with increasing ϵ. For example, the optimal value $T = 2$ for $\rho = 3$ lognormal distribution case given $\epsilon = 0$, gradually increases to the optimal value $T = 8$ when $\epsilon = 1$. The possible explanation is similar with that of Fig. 5.28.

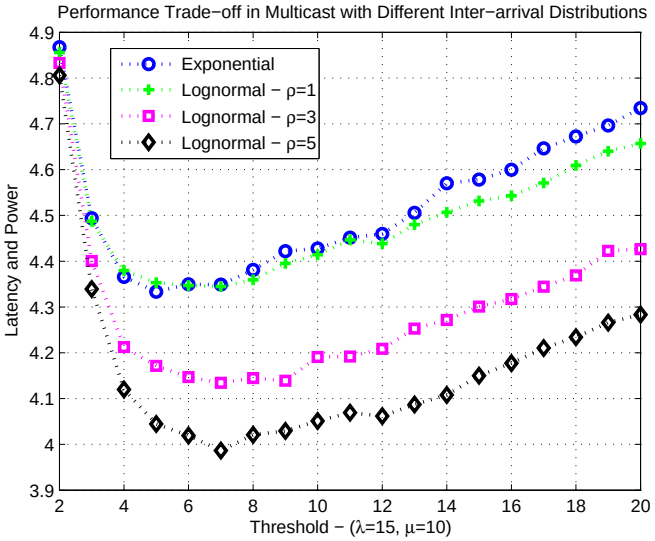


FIGURE 5.29: Latency and power trade-offs with multicast threshold T under different degrees of aggregation for lognormal distributed inter-arrival time.

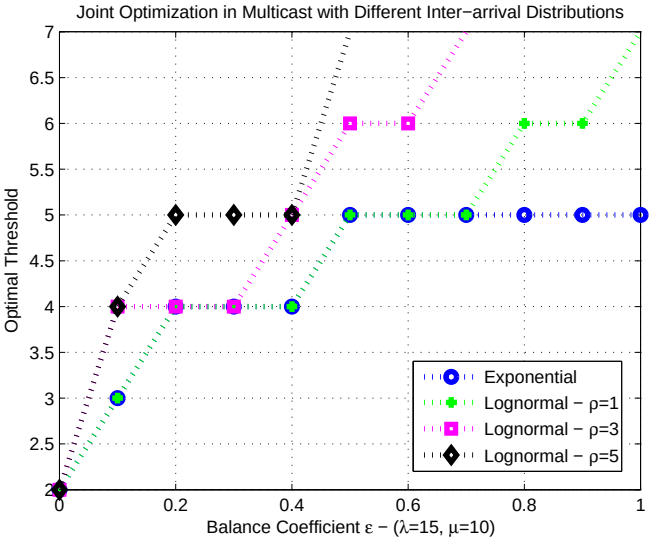


FIGURE 5.30: Optimal multicast thresholds for joint optimization of latency and power under different degrees of aggregation for lognormal distributed inter-arrival time.

5.3.4 Summary

In the above discussions, we analyzed the average latency of request and the average power consumption of BS when the content request temporally follows Poisson arrivals or bursty arrivals in the proposed unicast/multicast hybrid mechanism. Firstly, under the traditional Poisson arrivals, we use the Markov chain model to describe the proposed service process mathematically, and caculate the corresponding limit probabilities, average waiting time of each state in the model and the overall average latency in the theoretical derivation and simulation verification. Secondly, for the bursty arrival case, since there is no closed-form solution in theory, we use a numerical simulation method to compare the average latency, the multicast probability and the average power consumption for different inter-arrival distributions. Finally, in order to comprehensively balance the performance of latency and power consumption, we have jointly optimized them in Poisson and bursty arrival scenarios to find the optimal multicast thresholds minimizing the weighted sum of these two metrics. According to the results, we found that the adoption of the unicast/multicast hybrid service can not only solve the problem that the average latency of user request increases indefinitely with the arrival rate under congestion, but also being more effective for the practical bursty arrivals (less latency compare to Poisson arrivals with the same arrival rate). In addition, we proposed a new indicator ρ (the standard deviation of inter-arrival times divided by the corresponding mean) that describes the aggregation of content requests temporally, and found that higher ρ results in smaller latency and power consumption on average given arrival rates are equal, which further confirms the effectiveness of the hybrid strategy. Finally, through the joint optimization of latency and power consumption, we found that for different request arrival modes, different multicast thresholds should be chosen to minimize the overall performance of the hybrid strategy.

5.4 Intelligent SDN Architecture for Smart Caching and Dynamic Multicast

In the above two sections, we introduced the cooperative caching strategy for access networks based on the spatial clustering of BS and the unicast/multicast hybrid strategy based on the temporal aggregation effect of content requests. From the performance analysis of the two service strategies, we can see that the cooperative caching strategies in different scenarios (spatial distribution of BS, content preference skewness) have different optimal caching schemes; and the unicast/multicast strategy for different arrival modes (request arrival rate, degree of aggregation) is also not static. In order to always adopt the optimal strategy in real time, we need to transform the proposed two service solutions into (i) a cooperative caching policy that can be intelligently distributed and (ii) a unicast/multicast strategy with dynamic threshold by tracking specific service scenario changes and user demand patterns. In order to discover service scenarios and

changes of user demands in real time, we propose an intelligent SDN-based centralized service framework in cellular networks [117]. By introducing an intelligent center, we collect required BS location information, user request records, and resource allocation information and take advantages of corresponding algorithm dynamically to calculate the optimal cache probability on each BS and the optimal multicast threshold for different content. The overall service architecture envisaged is shown in Fig. 5.31.

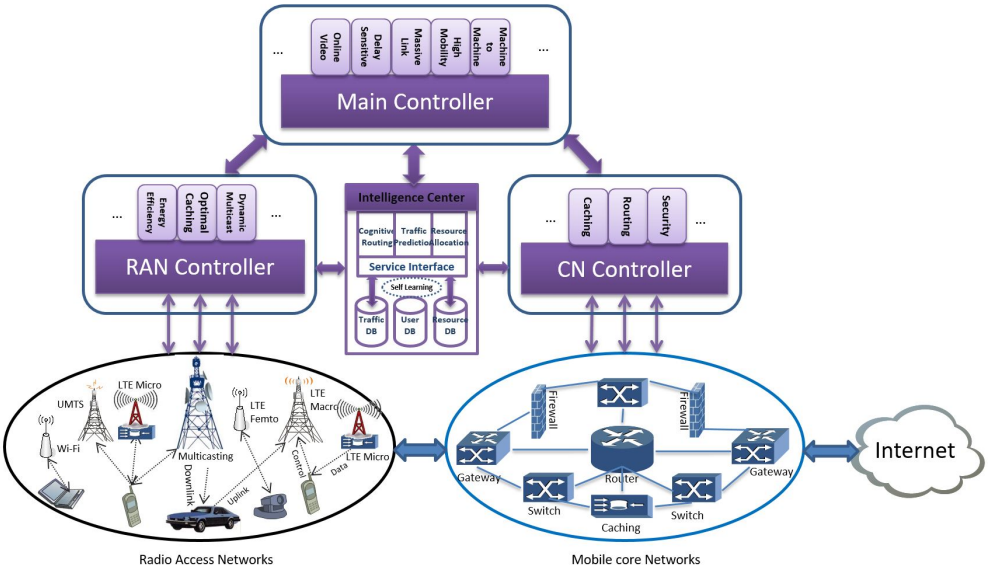


FIGURE 5.31: Intelligent SDN based heterogeneous cellular network architecture for efficient service strategy.

Specifically, for the cooperative caching strategy in BS cluster, the RAN controller first collects the user's content request information, the location and bandwidth resource information of each BS, and these data records are further transferred to the intelligent center. Within the intelligent center, spatial aggregation analysis of BSs and preference analysis of user content requests are performed through configurable functions or algorithms. For example, according to the clustering algorithm, all BSs can be divided into different cooperative clusters. Then, the popular contents in the area are identified by the content request records in different clusters and the popularity of these popular contents is calculated. In this way, by combining the clustering result of the BS, the popularity of the popular contents and the available bandwidth resources of BSs in the cluster, the caching probability for each content in each cluster can be calculated through the cache optimization algorithm provided in the above section. This decision process can be implemented based on the powerful computing and storage capabilities of the intelligent center and the results obtained will be sent to each RAN controller, and the controller will transfer corresponding buffer instructions to each BS. In this way, through the distributed data collection, the method of centralized data processing can fully utilize the practical value of the large amount of traffic data records, thus providing the possibility for intelligent and efficient caching.

On the other hand, for the proposed unicast/multicast hybrid strategy, in order to dynamically implement the optimized multicast threshold to jointly reduce the average latency of requests and the average power consumption of BS, it is also necessary to support the real data collection to accurately predict the popular contents. Based on the real data collected, the arrival rate and degree of aggregation of content requests can be estimated dynamically in advance. After that, corresponding multicast thresholds will be set for different request contents in different time slots. Similarly, the RAN controller may collect all access user's content request records through the BS within its coverage area, including content identification, requesting time, and delay requirements. Then, all the data are delivered to the intelligent center. After receiving all these information, the intelligence center can make timely traffic prediction for all popular contents, including the arrival rate and the degree of aggregation, with the historical data already existing in the database. Together with the balance coefficient between the average latency of request and the average power consumption of BS, an optimal multicast threshold can be set for each content, and the corresponding result is returned to the RAN controller. After receiving the instruction, the controller can deliver the corresponding multicast threshold for each content to the corresponding BS, thereby implementing a dynamic unicast/multicast service strategy.

By introducing such an intelligent centralized processing unit, the cellular networks can trace the timely demand variation more quickly, and therefore simultaneously satisfy the operational requirements of the entire network and service quality of all users by deploying flexible and efficient algorithms. As two efficient service strategies rooted in the aggregation of user demands, cache and multicast are in dire need of the support of large-scale real-time data, in order to achieve intelligent and dynamic service attributes. Therefore, the cooperative caching and dynamic multicast strategy proposed in this chapter emerge as two representative use cases in this intelligent SDN-based centralized service framework in cellular networks.

5.5 Conclusion and Discussion

In this chapter, we proceeded from the clustering nature of cellular networks obtained from real data in Chapter 4, and proposed a cooperative caching strategy for RAN based on the spatial aggregation of BSs and a unicast/multicast hybrid strategy based on the temporal aggregation of content requests.

In the cooperation-based probabilistic caching strategy, the BS can achieve the sharing of cache contents through the transmission bandwidth between different nodes within one cluster. At the same time, this strategy increases the total amount of cached content in the entire cooperative cluster, thereby shortening the average path length of the content delivery of users, thus reducing the average delay of users and data congestion in the backbone network. Meanwhile, we also

found that the spatially aggregated BS distribution is more suitable for the cooperative cache strategy than the homogeneous BS deployment. This is because the corresponding overall performance can increase beyond the linearity growth trend as the number of cooperative BSs within the cluster increases.

In addition, in the proposed unicast/multicast hybrid transmission strategy, we theoretically analyzed and compared the average latency of request and the average power consumption of BS under two types of content request modes, i.e., Poisson arrivals and bursty arrivals. The results showed that, unlike the traditional unicast strategy, the introduction of the multicast mechanism can perfectly solve the problem that the user latency increases rapidly when the request arrival rate exceeds the service capacity of the BS. At the same time, it can also reduce the BS's average power consumption for serving all user requests. On the other hand, for bursty arrival patterns, we raised a new coefficient ρ that can describe the degree of the temporal aggregation of content requests flow, defined as the quotient of the standard deviation of inter-arrival time and corresponding mean value. Through simulation verification, we found that, with the same content request arrival rate, the larger the aggregation coefficient, the greater the multicast probability of service queue. Therefore, it leads to the smaller average latency of requests and the smaller average power consumption of BS, which highlights the significant benefit of this hybrid strategy for contant requests under bursty arrivals. Besides, for different arrival patterns, in order to comprehensively consider both the average latency and the average power consumption, we propose a joint optimization scheme based on a balance coefficient ϵ. In this joint optimization, an optimal multicast threshold is selected to minimize the weighted sum of latency and power consumption for different arrival rates and degree of temporal aggregation.

Finally, because of the rapid dynamics of user content preferences and the bursty nature of user requests, in order to accurately complete these two high-efficiency service strategies in real time, we proposed an intelligent SDN-based centralized service framework in cellular networks. Through the collection of a large amount of real-time records, the efficient algorithms deployed on the intelligence center can provide timely prediction and solution for different tasks. As two use cases, the proposed architecture provides computing and storage resources for the cooperative cache strategy and the unicast/multicast hybrid transmission strategy, to achieve intelligent and dynamic service attributes.

Chapter 6

Conclusion and Future Works

Contents

6.1 Conclusion

This thesis focused on the unveiling of clustering nature in cellular networks, and evaluate their potential impacts on the service performance. Specifically, the clustering nature includes traffic pattern, infrastructure deployment and content requests on temporal, spatial and content dimension, respectively. To utilize those clustering nature, based on the spatial aggregation of BS deployment and the temporal burstiness of content requests, we stepped forward to propose cooperative caching and dynamic multicast in cellular networks which significantly improves the service performance according to corresponding results.

After introducing the background and motivation of our work in Chapter 1, we gave a comprehensive review of the state-of-the-art real data measurement in Chapter 2 which not only sheds light on the importance of real data analysis, but also paves the way for its reasonable usage to improve the service performance of cellular networks. Following the motivation and related works in this literature, we tried to uncover the clustering nature of cellular networks based on collected measurement from telecommunication operator in China, which can be found in Chapter 3 and Chapter 4. Based on the empirical results, we proposed cooperative caching and dynamic multicast to utilize the clustering phenomenon as presented in Chapter 5. In detail, the corresponding conclusion can be summarized as follows.

- Based on the review in Chapter 2, we concluded that there exhibits a periodic pattern of the temporal traffic assumption for large coverage area in cellular networks, while

for single cell, a heavy-tailed distribution is widespread across the temporal and spatial characterization of traffic consumption. Furthermore, the imbalance phenomenon emerges more significantly in the call duration, request arrivals and content preference of mobile users, where mathematical description and adequate real data are necessary for more accurate characterization.

- Based on a large amount of real data collected from on-operating Chinese cellular networks, we conducted a large-scale identification on spatial modeling of BS in Chapter 3. According to the fitting results, we verified the inaccuracy of PPP's usage for BS locations, and uncovered the clustering nature of BS deployment in cellular networks. Although the two typical clustering models (MCP and TCP) have improved the modeling accuracy but are still not qualified to accurately reproduce the practical BSs deployment, which leads to the spatial density characterization of BS in next chapter.

- In Chapter 4, we characterized the density of BS deployment and traffic demand, in both spatial and temporal dimensions. In accordance with the heavy-tailed phenomenons in Chapter 2, we found that the α-Stable distribution is the most accurate model for the spatial densities of BSs and traffic consumption, between which a linear dependence was revealed through real data examinations. Moreover, the accuracies of power-law and lognormal distributions for the packet length and inter-arrival time of user requests were verified, respectively, which convincingly leads to the α-Stable distribution of temporally aggregated traffic volume on BS level.

- To make benefit from the findings of previous chapters, we proposed a cooperative caching strategy in RAN based on the spatial aggregation of BSs and a dynamic unicast/multicast strategy based on the temporal aggregation of content requests in Chapter 5. According to the theoretical and simulation results, we found that the proposed 'Caching as a Cluster' strategy can significantly reduce the average delay of users especially in the inhomogeneous BS deployment scenario. Besides, the dynamic unicast/multicast strategy can not only reduce the average latency of content requests but also diminishing the average power consumption of BSs especially under the bursty request arrival patterns.

- To implement the massive real data analyses and dynamical serving mechanism, we proposed an intelligent SDN-based centralized architecture within cellular networks in Chapter 5. With the introduction of an intelligence center, the brand new architecture is able to trace the demand variations in real time, thus simultaneously satisfy the operational requirements of the entire network and QoE requirements of all users by deploying flexible and efficient algorithms upon it.

Conclusively, in this thesis, we uncover the clustering nature of cellular networks in different dimensions, and proposed corresponding service strategies to tackle the clustering challenge and utilize them for efficiency improvement.

6.2 Future Works

In our work, although the real data was already sufficient for the identification process in quantity, they were all collected from telecommunication operators in China, thus it's not convincible to apply the conclusion worldwide. In order to conduct a comprehensive study and obtain a more accurate model for the traffic characterization, we need real data from more countries with different geographical and population properties. On the other hand, the traffic records adopted in this work were collected on the BS level at most, which lacks more fine-grained measurement on the mobile user level. Although there should be a privacy consideration on the data collection procedure, anonymous records on user level are welcomed for more accurate characterization of clustering nature in cellular networks.

In Chapter 3 and Chapter 4, we made a comprehensive description of the clustering nature in cellular networks based on a large amount of real data. Although it's adequate and convincible to uncover this phenomenon by empirical results, it's also indispensable to characterize the clustering effect on different dimensions by mathematical description which is lacking in this thesis. More intuitive and incisive theoretical characterization of the clustering nature in cellular networks can be considered as a promising future work.

In Chapter 5, we separately considered the caching and multicast technics to utilize the clustering nature of cellualr networks. Actually, the caching strategy makes use of the spatial clustering of BSs and the content preference of mobile users, and the multicast technic utilizes the temporal aggregation of user requests and the broadcasting nature of wireless signal. In a word, the caching strategy is mainly to optimize and improve the part of the wired transmission, while multicast technic is focused on the wireless transmission part of the content transmission. Therefore, there is a possibility to combine these two technics to fully exploit the optimization potential of the clustering nature in different dimensions.

In the intelligent SDN-based architecture proposed in Chapter 5, we mainly focused on the ability of the new structure to enable the real data analysis and service policy, but doesn't consider the enabling technologies inside it. In our point of view, the algorithm running in the intelligence center is of utmost importance to implement the whole procedure which inevitably involves big data technics and machine learning algorithms. In order to reasonably and efficiently utilize the real data in cellular networks, some customized technics and algorithms are

essential for the effectiveness of the new proposed intelligent SDN-based architecture which can be considered as another promising direction in this literature.

Appendix A

List of Publications

A.1 Journal Papers

1. **Yifan Zhou**, Zhifeng Zhao, Rongpeng Li, Honggang Zhang and Yves Louet, "Cooperation-Based Probabilistic Caching Strategy in Clustered Cellular Networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2029-2032, Sept. 2017.

2. **Yifan Zhou**, Zhifeng Zhao and Honggang Zhang, "Towards 5G: Heterogeneous Cellular Network Architecture Design Based on Intelligent SDN Paradigm," *Telecommunications Science*, vol. 32, no. 6, pp. 28-35, Jun. 2016.

3. **Yifan Zhou**, Zhifeng Zhao, Yves Louet, Qianlan Ying, Rongpeng Li, Xuan Zhou, Xianfu Chen and Honggang Zhang, "Large-scale Spatial Distribution Identification of Base Stations in Cellular Networks," *IEEE Access*, vol. 3, pp. 2987-2999, Dec. 2015.

4. **Yifan Zhou**, Rongpeng Li, Zhifeng Zhao, Xuan Zhou, and Honggang Zhang, "On the alpha-Stable Distribution of Base Stations in Cellular Networks," *IEEE Commun. Lett.*, vol.19, no.10, pp.1750-1753, Aug. 2015.

5. Luca Chiaraviglio, Francesca Cuomo, Maurizio Maisto, Andrea Gigli, Josip Lorincz, **Yifan Zhou**, Zhifeng Zhao, Chen Qi, Honggang Zhang, "What is the Best Spatial Distribution to Model Base Station Density? A Deep Dive in Two European Mobile Networks," *IEEE Access*, vol. 4, pp. 1434-1443, Apr. 2016.

6. Rongpeng Li, Zhifeng Zhao, Chen Qi, Xuan Zhou, **Yifan Zhou**, and Honggang Zhang, "Understanding the Traffic Nature of Mobile Instantaneous Messaging in Cellular Networks: A Revisiting to α-Stable Models," *IEEE Access*, vol. 3, pp. 1416-1422, Aug. 2015.

7. Xuan Zhou, Zhifeng Zhao, Rongpeng Li, **Yifan Zhou**, Tao Chen, Zhisheng Niu and Honggang Zhang, "Towards 5G: When Explosive Bursts Meet Soft Cloud," *IEEE Netw.*, vol. 28, no. 6, pp. 12–17, Nov. 2014.

8. Xuan Zhou, Zhifeng Zhao, Rongpeng Li, **Yifan Zhou**, Jacques Palicot and Honggang Zhang, "Understanding the Nature of Social Mobile Instant Messaging in Cellular Networks," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 389 – 392, Mar. 2014.

9. Xuan Zhou, Zhifeng Zhao, Rongpeng Li, **Yifan Zhou**, Jacques Palicot and Honggang Zhang, "Human Mobility Patterns in Cellular Networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1877-1880, Sep. 2013.

A.2 International Conference Papers

1. **Yifan Zhou**, Yves Louet and Honggang Zhang, "Probabilistic Caching Strategy in Collaborative Small Cell Networks," in *Proc. 32th URSI GASS*, Montreal, Canada, Aug. 2017.

2. **Yifan Zhou**, Zhifeng Zhao, Qianlan Ying, Rongpeng Li, Xuan Zhou and Honggang Zhang, "Two-tier Spatial Modeling of Base Stations in Cellular Networks," in *Proc. IEEE PIMRC 2014*, Washington DC, USA, Sep. 2014.

3. Meng Li, Zhifeng Zhao, **Yifan Zhou**, Xianfu Chen and Honggang Zhang, "On the Dependence Between Base Stations Deployment and Traffic Spatial Distribution in Cellular Networks," in *Proc. International Conference on Telecommunications 2016*, Thessaloniki, Greece, May 2016.

4. Luca Chiaraviglio, Francesca Cuomo, Andrea Gigli, Maurizio Maisto, **Yifan Zhou**, Zhifeng Zhao, Honggang Zhang, "A Reality Check of Base Station Spatial Distribution in Mobile Networks," in *IEEE INFOCOM 2016 (Poster)*, San Francisco, Apr. 2016.

5. Xianfu Chen, Celimuge Wu, **Yifan Zhou**, Honggang Zhang, "A learning approach for traffic offloading in stochastic heterogeneous cellular networks," in *Proc. IEEE ICC 2015*, London, UK, Jun. 2015.

A.3 Seminars or Presentations

1. **Yifan Zhou**, Large-scale Real Data Based Spatial Analysis of Base Stations, The First Worshop on Big Data in Wireless Networks, Oct. 2014, Hefei.

2. **Yifan Zhou**, A Revisiting to Queueing Theory for Mobile Instant Messaging with Keep-Alive Mechanism in Cellular Networks,ICC 2017 Technical Symposia, May 2017, Paris, France.

3. **Yifan Zhou**, Clustering Nature of BS Deployment and Traffic Demand in Cellular Networks, SCEE Seminar, Sep. 2017, Rennes, France.

List of Figures

List of Tables

Bibliography

[1] 3GPP. GERAN study on mobile data applications. Technical Report 3GPP TR 43.802, May 2011.

[2] H. Ahlehagh and S. Dey. Video caching in radio access network: Impact on delay and capacity. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 2276–2281. IEEE, 2012.

[3] S. Almeida, J. Queijo, and L. M. Correia. Spatial and temporal traffic distribution models for gsm. In *Vehicular Technology Conference, 1999. VTC 1999-Fall. IEEE VTS 50th*, volume 1, pages 131–135. IEEE, 1999.

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

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

[6] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera. Multicasting over emerging 5g networks: Challenges and perspectives. *IEEE Network*, 31(2):80–89, 2017.

[7] A. Baddeley and R. Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[9] E. Bastug, M. Bennis, and M. Debbah. Living on the edge: The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, 52(8):82–89, 2014.

[10] B. Blaszczyszyn and A. Giovanidis. Optimal geographic caching in cellular networks. In *Proceedings of ICC, IEEE*, pages 3358–3363, 2015.

[11] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM*, volume 1, pages 126–134. IEEE, 1999.

[12] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo. Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1430–1441, 2012.

[13] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[14] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, 2007.

[15] V. Chandrasekhar and J. G. Andrews. Spectrum allocation in tiered cellular networks. *IEEE Trans. Commun.*, 57(10):3059–3068, 2009.

[16] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo. Probabilistic small-cell caching: Performance analysis and optimization. *IEEE Transactions on Vehicular Technology*, 66(5):4341–4354, 2017.

[17] L. Chenand, W. Chen, B. Wang, X. Zhang, H. Chen, and D. Yang. System-level simulation methodology and platform for mobile cellular systems. *IEEE Communications Magazine*, 49(7), 2011.

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

[19] I. Chih-Lin, J. Huang, R. Duan, C. Cui, J. X. Jiang, and L. Li. Recent progress on c-ran centralization and cloudification. *IEEE Access*, 2:1030–1039, 2014.

[20] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, 2013.

[21] S. Cho and W. Choi. Energy-efficient repulsive cell activation for heterogeneous cellular networks. *IEEE J. Sel. Area. Commun.*, 31(5):870–882, 2013.

[22] Y. J. Chun, M. O. Hasna, and A. Ghrayeb. Modeling heterogeneous cellular networks interference using poisson cluster processes. *IEEE Journal on Selected Areas in Communications*, 33(10):2182–2195, 2015.

[23] Cisco. Cisco visual network index: Global mobile data traffic forecast update, 2015 - 2020 white paper. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2016. [Online].

[24] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, Dec. 1997.

[25] N. Deng, W. Zhou, and M. Haenggi. The Ginibre point process as a model for wireless networks with repulsion. *arXiv preprint arXiv:1401.3677*, 2014.

[26] N. Deng, W. Zhou, and M. Haenggi. A heterogeneous cellular network model with inter-tier dependence. In *Proc. IEEE GLOBECOM2014*, Austin, TX, 2014.

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

[28] J. Ding, X. Liu, Y. Li, D. Wu, D. Jin, and S. Chen. Measurement-driven capability modeling for mobile network in large-scale urban environment. In *Mobile Ad Hoc and Sensor Systems (MASS), 2016 IEEE 13th International Conference on*, pages 92–100. IEEE, 2016.

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

[30] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck. To cache or not to cache: The 3g case. *IEEE Internet Computing*, 15(2):27–34, 2011.

[31] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck. Over the top video: the gorilla in cellular networks. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 127–136. ACM, 2011.

[32] J. Erman and K. K. Ramakrishnan. Understanding the super-sized traffic of the super bowl. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 353–360. ACM, 2013.

[33] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. ACM SIGCOMM 1999*, Massachusetts, USA, 1999.

[34] J. R. Gallardo, D. Makrakis, and L. Orozco-Barbosa. Use of α-stable self-similar stochastic processes for modeling traffic in broadband networks. *Performance Evaluation*, 40(1):71–98, 2000.

[35] R. K. Ganti and M. Haenggi. Interference and outage in clustered wireless ad hoc networks. *IEEE Trans. Inf. Theory*, 55(9):4067–4086, 2009.

[36] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28. ACM, 2007.

[37] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[38] U. Gotzner and R. Rathgeber. Spatial traffic distribution in cellular networks. In *Vehicular Technology Conference, 1998. VTC 98. 48th IEEE*, volume 3, pages 1994–1998. IEEE, 1998.

[39] Y. Guan. A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476):1502–1512, 2006.

[40] A. Guo and M. Haenggi. Spatial stochastic models and metrics for the structure of base stations in cellular networks. *IEEE Trans. Wireless. Commun.*, 12:5800–5812, Nov. 2013.

[41] J. Guo, F. Liu, and Z. Zhu. Estimate the call duration distribution parameters in gsm system based on kl divergence method. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 2988–2991. IEEE, 2007.

[42] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE J. Sel. Area. Commun.*, 27(7):1029–1046, Sep. 2009.

[43] F. Hartung, U. Horn, J. Huschke, M. Kampmann, and T. Lohmar. Mbms ip multicast/broadcast in 3g networks. *International Journal of Digital Multimedia Broadcasting*, 2009, 2009.

[44] K. Huang and V. K. Lau. Enabling wireless power transfer in cellular networks: Architecture, modeling and deployment. *IEEE Transactions on Wireless Communications*, 13(2):902–912, 2014.

[45] M. Ji, G. Caire, and A. F. Molisch. Wireless device-to-device caching networks: Basic principles and system performance. *IEEE Journal on Selected Areas in Communications*, 34(1):176–189, 2016.

[46] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson, S. Sen, S. Venkataraman, and Z.-L. Zhang. Characterizing data usage patterns in a large cellular network. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, pages 7–12. ACM, 2012.

[47] A. N. Kolmogorov, K. L. Chung, and B. V. Gnedenko. *Limit distributions for sums of independent random variables*. Addison-Wesley, Reading, Mass, rev. ed. edition, 1968.

[48] M. Laner, P. Svoboda, S. Schwarz, and M. Rupp. Users in cells: A data traffic analysis. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 3063–3068. IEEE, 2012.

[49] C.-H. Lee, C.-Y. Shih, and Y.-S. Chen. Stochastic geometry based models for modeling cellular networks in urban areas. *Wireless Networks*, 19(6):1063–1072, Aug. 2013.

[50] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang. Spatial modeling of the traffic density in cellular networks. *IEEE Wireless Communications*, 21(1):80–88, 2014.

[51] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[52] K. Li, C. Yang, Z. Chen, and M. Tao. Optimization and analysis of probabilistic caching in n-tier heterogeneous networks. *arXiv preprint:1612.04030*, 2016.

[53] M. Li, Z. Zhao, Y. Zhou, X. Chen, and H. Zhang. On the dependence between base stations deployment and traffic spatial distribution in cellular networks. In *Telecommunications (ICT), 2016 23rd International Conference on*, pages 1–5. IEEE, 2016.

[54] R. Li, Z. Zhao, C. Qi, X. Zhou, Y. Zhou, and H. Zhang. Understanding the traffic nature of mobile instantaneous messaging in cellular networks: A revisiting to α-stable models. *IEEE Access*, 3:1416–1422, Sept. 2015.

[55] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang. The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice. *IEEE Communications Magazine*, 52(6):234–240, 2014.

[56] Y. Li, F. Baccelli, H. S. Dhillon, and J. G. Andrews. Fitting determinantal point processes to macro base station deployments. In *Proc. IEEE GLOBECOM2014*, Austin, TX, Dec. 2014.

[57] D. Liu, B. Chen, C. Yang, and A. F. Molisch. Caching at the wireless edge: Design aspects, challenges, and future directions. *IEEE Communications Magazine*, 54(9):22–28, 2016.

[58] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 90–102. ACM, 2009.

[59] J. H. McCulloch. Simple consistent estimators of stable distribution parameters. *Communications in Statistics-Simulation and Computation*, 15(4):1109–1136, 1986.

[60] J. H. McCulloch. Simple consistent estimators of stable distribution parameters. *Commun. Stat. Simulat.*, 15(4):1109–1136, Jan. 1986.

[61] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean. Single frequency-based device-to-device-enhanced video delivery for evolved multimedia broadcast and multicast services. *IEEE Transactions on Broadcasting*, 61(2):263–278, 2015.

[62] M. Mirahsan, R. Schoenen, and H. Yanikomeroglu. Hethetnets: Heterogeneous traffic distribution in heterogeneous wireless cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(10):2252–2265, 2015.

[63] J. Moller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton, 2003.

[64] D. Naboulsi. *Analysis and exploitation of mobile traffic datasets*. PhD thesis, Lyon, INSA, 2015.

[65] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[66] J. Neyman and E. Scott. Processes of clustering and applications. *Stochastic Point Processes*, pages 646–681, 1972.

[67] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute. Measurement-driven mobile data traffic modeling in a large metropolitan area. In *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*, pages 230–235. IEEE, 2015.

[68] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamics in cellular data networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 882–890. IEEE, 2011.

[69] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas. Exploiting caching and multicast for 5g wireless networks. *IEEE Transactions on Wireless Communications*, 15(4):2995–3007, 2016.

[70] B. A. Ramanan, L. M. Drabeck, M. Haner, N. Nithi, T. E. Klein, and C. Sawkar. Cacheability analysis of http traffic in an operational lte network. In *Wireless Telecommunications Symposium (WTS), 2013*, pages 1–8. IEEE, 2013.

[71] T. S. Rappaport et al. *Wireless communications: principles and practice*, volume 2. Prentice Hall PTR New Jersey, 1996.

[72] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[73] J. Riihijarvi and P. Mahonen. Modeling spatial structure of wireless communication networks. In *Proc. IEEE INFOCOM2010*, San Diego, CA, Mar. 2010.

[74] B. Ripley. Modelling spatial patterns. *J. Royal Statistical Society, Series B*, 39:172–212, 1977.

[75] B. D. Ripley. *Statistical Inference for Spatial Processes*. Cambridge University Press, 1991.

[76] S. Rosen, H. Luo, Q. A. Chen, Z. M. Mao, J. Hui, A. Drake, and K. Lau. Discovering fine-grained rrc state dynamics and performance impacts in cellular networks. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 177–188. ACM, 2014.

[77] S. M. Ross. *Introduction to probability models*. Academic press, 2014.

[78] G. Samorodnitsky. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall/CRC, New York, June 1994.

[79] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang. Understanding the impact of network dynamics on mobile video user engagement. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 367–379. ACM, 2014.

[80] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang. A first look at cellular network performance during crowded events. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 17–28. ACM, 2013.

[81] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang. Characterizing and optimizing cellular network performance during crowded events. *IEEE/ACM Transactions on Networking*, 24(3):1308–1321, 2016.

[82] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. Characterizing geospatial dynamics of application usage in a 3g cellular data network. In *INFOCOM, 2012 Proceedings IEEE*, pages 1341–1349. IEEE, 2012.

[83] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 305–316. ACM, 2011.

[84] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Transactions on Information Theory*, 59(12):8402–8413, 2013.

[85] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[86] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp, and H. Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243, 2013.

[87] W. L. Tan, F. Lam, and W. C. Lau. An empirical study on the capacity and performance of 3g networks. *IEEE Transactions on Mobile Computing*, 7(6):737–750, 2008.

[88] A. Tassi, I. Chatzigeorgiou, and D. Vukobratović. Resource-allocation frameworks for network-coded layered multimedia multicast services. *IEEE Journal on Selected Areas in Communications*, 33(2):141–155, 2015.

[89] D. B. Taylor, H. S. Dhillon, T. D. Novlan, and J. G. Andrews. Pairwise interaction processes for modeling cellular network topology. In *Proc. IEEE GLOBECOM2012*, Anaheim, CA, Dec. 2012.

[90] R. Trasarti, A.-M. Olteanu-Raimond, M. Nanni, T. Couronné, B. Furletti, F. Giannotti, Z. Smoreda, and C. Ziemlicki. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, 39(3):347–362, 2015.

[91] K. Tutschku and P. Tran-Gia. Spatial traffic estimation and characterization for mobile communication network design. *IEEE Journal on selected areas in communications*, 16(5):804–811, 1998.

[92] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin. Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks. In *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile computing and online Social neTworking*, pages 19–24. ACM, 2015.

[93] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 225–238. ACM, 2015.

[94] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung. Cache in the air: Exploiting content caching and delivery techniques for 5g systems. *IEEE Communications Magazine*, 52(2):131–139, 2014.

[95] C. Williamson, E. Halepovic, H. Sun, and Y. Wu. Characterization of cdma2000 cellular data network traffic. In *Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on*, pages Z000–719. IEEE, 2005.

[96] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary users in cellular networks: A large-scale measurement study. In *New frontiers in dynamic spectrum access networks, 2008. DySPAN 2008. 3rd IEEE symposium on*, pages 1–11. IEEE, 2008.

[97] P. E. Wirth. The role of teletraffic modeling in the new communications paradigms. *IEEE Communications Magazine*, 35(8):86–92, 1997.

[98] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park. Comparison of caching strategies in modern cellular backhaul networks. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 319–332. ACM, 2013.

[99] L. Wu, Y. Zhong, and W. Zhang. Spatial statistical modeling for heterogeneous cellular networks–an empirical study. In *Proc. IEEE VTC2014-Spring*, Seoul, Korea, May 2014.

[100] F. H. Z. Xavier, L. M. Silveira, J. M. d. Almeida, A. Ziviani, C. H. S. Malab, and H. T. Marques-Neto. Analyzing the workload dynamics of a mobile phone network in large scale events. In *Proceedings of the first workshop on Urban networking*, pages 37–42. ACM, 2012.

[101] Z. Xiao, L. Guo, and J. Tracey. Understanding instant messaging traffic characteristics. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, pages 51–51. IEEE, 2007.

[102] G. Xiaohu, Z. Guangxi, and Z. Yaoting. On the testing for alpha-stable distributions of network traffic. *Computer Communications*, 27(5):447–457, Mar. 2004.

[103] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li. Big data driven mobile traffic understanding and forecasting: A time series approach. *IEEE Transactions on Services Computing*, 9(5):796–805, 2016.

[104] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao. Cellular data network infrastructure characterization and implication on mobile content placement. In *Proceedings of the SIGMETRICS, ACM*, pages 317–328, 2011.

[105] C. Yang, Y. Yao, Z. Chen, and B. Xia. Analysis on cache-enabled wireless heterogeneous networks. *IEEE Transactions on Wireless Communications*, 15(1):131–145, 2016.

[106] P. Zerfos, X. Meng, S. H. Wong, V. Samanta, and S. Lu. A study of the short message service of a nationwide cellular network. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 263–268. ACM, 2006.

[107] J. Zhang, W. Wang, X. Zhang, Y. Huang, Z. Su, and Z. Liu. Base stations from current mobile cellular networks: Measurement, spatial modeling and analysis. In *Wireless Communications and Networking Conference Workshops (WCNCW), 2013 IEEE*, pages 1–5. IEEE, 2013.

[108] Y. Zhang and A. Årvidsson. Understanding the characteristics of cellular data traffic. *ACM SIGCOMM Computer Communication Review*, 42(4):461–466, 2012.

[109] S. Zhou, D. Lee, B. Leng, X. Zhou, H. Zhang, and Z. Niu. On the spatial distribution of base stations and its relation to the traffic density in cellular networks. *IEEE Access*, 3:998–1010, 2015.

[110] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang. Human mobility patterns in cellular networks. *IEEE Communications Letters*, 17(10):1877–1880, 2013.

[111] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang. Understanding the nature of social mobile instant messaging in cellular networks. *IEEE Communications Letters*, 18(3):389–392, 2014.

[112] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang. The predictability of cellular networks traffic. In *Communications and Information Technologies (ISCIT), 2012 International Symposium on*, pages 973–978. IEEE, 2012.

[113] Y. Zhou, R. Li, Z. Zhao, X. Zhou, and H. Zhang. On the α-stable distribution of base stations in cellular networks. *IEEE Commun. Letters*, 19(10):1750–1753, Aug. 2015.

[114] Y. Zhou, Z. Zhao, R. Li, H. Zhang, and Y. Louet. Cooperation-based probabilistic caching strategy in clustered cellular networks. *IEEE Communications Letters*, 21(9):2029–2032, 2017.

[115] Y. Zhou, Z. Zhao, Y. Louët, Q. Ying, R. Li, X. Zhou, X. Chen, and H. Zhang. Large-scale spatial distribution identification of base stations in cellular networks. *IEEE access*, 3:2987–2999, 2015.

[116] Y. Zhou, Z. Zhao, Q. Ying, R. Li, X. Zhou, and H. Zhang. Two-tier spatial modeling of base stations in cellular networks. In *Proc. IEEE PIMRC2014*, Washington DC, USA, Sep. 2014.

[117] Y. Zhou, Z. Zhao, Zhifeng, and Z. Honggang. Towards 5g: heterogeneous cellular network architecture design based on intelligent sdn paradigm. *Telecommunications Science*, 32(6):28, 2016.