UNIVERSITÀ DI PADOVA FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

SCUOLA DI DOTTORATO IN INGEGNERIA DELL'INFORMAZIONE

INDIRIZZO IN SCIENZA E TECNOLOGIA DELL'INFORMAZIONE

XXVIII Ciclo

Context-Aware Optimization in Heterogeneous Networks: Handover and Caching Strategies

Dottoranda

IRENE PAPPALARDO

Supervisore:

Chiar.mo Prof. Michele Zorzi

Direttore della Scuola:

Chiar.mo Prof. Matteo Bertocco

Anno Accademico 2015/2016

Abstract

The pervasive and progressive distribution of the 4G network is associated with the rapid increase of new generation smartphones and LTE devices that are expected to guarantee a higher and higher performance and to provide an excellent quality of service to the users. After two months from the introduction of tablets to the market, 50 million mobile users were connected to the Internet through tablets, while it took four years to reach the same results with personal computers. Besides, users are more and more resource demanding and seek seamless connectivity everywhere. The need to face the consequent huge amount of data traffic produced and of resources required is posing a vast range of challenges that call for both an improved network infrastructure and new paradigms for resource management and optimization. The new frontier of next generation mobile systems relies on the concept of heterogeneous networks. These highly sophisticated systems accommodate multiple tiers of access nodes, representing thus a real break from the traditional network with a macrocell only topology. As a result, innovative flexible ways for resource management need to be invented, since the previous schemes are unsuitable or partially inadequate. This dissertation makes a step forward in that direction and proposes the novel exploitation of the so called *context information*, i.e., system parameters and other metrics related to the specific problem considered, to develop efficient resource optimization algorithms.

The issue of efficient context-aware resource management is considered in two scenarios, namely handover and caching optimization. In the former, the addressed challenge is the choice of two key parameters that regulate the offloading of traffic from the macro to the small cells, namely the time-to-trigger and the hysteresis margin. In the latter, instead, the goal is the optimal allocation of content in the caches of the base stations and of the users. Several parameters are considered as context in both cases, in particular the ones re-

lated to the conditions of the transmission channel, of the macro and small cells, and of the users, leading to extremely involved objective functions. In the two considered scenarios, a rigorous mathematical model that describes the system is presented and the consequent optimization framework is derived, leading to the minimization of the overall system cost or to the maximization of the users satisfaction.

In particular, concerning the optimization of the handover procedure, a general framework that analytically describes the process is presented, together with the characterization of the performance of a mobile user crossing a heterogeneous network, as a function of the users' mobility, of the power profile of the neighboring cells, of the handover parameters and of the traffic load of the different cells. Unlike many solutions in the literature that show only heuristic optimization methods, a rigorous Markov-based framework is used to model the handover process for the mobile user, and an optimal context-dependent algorithm is proposed. The mathematical model is validated by means of simulations, comparing the performance of the presented strategy with conventional handover optimization techniques in different scenarios. Finally, a general scheme to compute the performance upper bound of any handover algorithm proposed in the literature is investigated. The proposed solution is useful not only to determine the margin of improvement of existing handover schemes, but also to provide a comparative performance analysis among them.

Regarding the caching optimization framework, a novel system model is investigated, that comprises storage capability both at the small cells and at the mobile users. The requested content can be provided through a device-to-device communication from another peer node, or through a cellular downlink channel from the macro or the small cells. Users are hence encouraged to share their cached content, but they can establish a standard connection to the cell if the peer search fails. Within the considered heterogeneous scenario, the average system cost is derived in closed form, as a function of the user mobility pattern and of the content interest profile process. The derived optimization framework is implemented and significant performance gains are shown through simulation as compared to static caching policies. The proposed approach is original since an exhaustive analysis of strategies that exploit the opportunity of caching at the small cells and at the end devices was still missing. Moreover, in most other works users were often assumed to be static and to have the same interests, i.e., they are likely to request the same set of popular files,

whereas in practice different groups of users may have different preferences, which is explicitly accounted for in our model.

Sommario

La progressiva e pervasiva diffusione della nuova rete 4G si associa al sempre più diffuso utilizzo degli smartphone e tablet di nuova generazione e dei dispositivi LTE, in grado di garantire sempre maggiori prestazioni e di fornire servizi di altissima qualità agli utenti. Sono bastati solo ottanta giorni ai tablet per raggiungere i cinquanta milioni di utenti nel mondo. Ottanta giorni contro i quattro anni impiegati dai personal computer per arrivare al medesimo risultato. Inoltre, gli utenti sono sempre più esigenti e richiedono connettività senza interruzioni e ovunque. Per far fronte alla grandissima quantità di traffico dati prodotto e al conseguente aumento di risorse richieste è ovviamente necessario affrontare un'ampia gamma di sfide tecnologiche. Queste riguardano sia il miglioramento dell'infrastruttura di rete sia la progettazione di nuovi paradigmi per la gestione e l'ottimizzazione di risorse. La nuova frontiera dei sistemi radio mobili di prossima generazione si basa sul concetto di reti eterogenee. Quest'ultime sono sistemi altamente sofisticati in grado di ospitare diverse tipologie di nodi d'accesso, dalle macro celle alle celle di dimensioni inferiori con le quali gli operatori possono alleggerire il carico della macro cella, aumentando capacità e copertura. Per questo motivo la proliferazione delle celle secondarie, come pico e femto celle, decreta una vera e propria rottura dalla struttura di rete tradizionale, costituita da una topologia più semplice con sole macro celle. Nasce quindi il bisogno di inventare soluzioni originali e flessibili per la gestione delle risorse, soprattutto perchè gli schemi utilizzati in precedenza non sono più adeguati. Questa tesi si propone di fare un passo avanti in tale direzione e suggerisce uno studio basato sull'utilizzo innovativo delle cosiddette *informazioni di contesto* per sviluppare algoritmi efficienti di ottimizzazione delle risorse. Rientrano nelle informazioni di contesto qualsiasi parametro di sistema e metrica legata allo specifico problema considerato.

Lo studio della gestione di risorse basato sulla conoscenza di informazioni di contesto

è stato considerato in due diversi scenari, ovvero nell'ottimizzazione di tecniche di handover e nella progettazione di strategie di caching. Nel primo approccio, l'obiettivo è la scelta del valore ottimo di due parametri chiave, ovvero il tempo di trigger e il margine di isteresi, che regolano il trasferimento di traffico dalla macrocella a celle più piccole e viceversa. L'obiettivo nel secondo approccio, invece, è l'allocazione ottima dei contenuti nelle cache delle stazioni base e degli utenti.

In entrambi gli scenari considerati, viene presentato un modello matematico rigoroso che descrive il sistema e successivamente viene studiata l'ottimizzazione della funzione obiettivo scelta. Inoltre, si considerano diversi parametri come informazione di contesto, ad esempio relativi alle condizioni della macro cella, delle celle secondarie e degli utenti, e al profilo di trasmissione del canale.

In particolare, per quanto riguarda l'ottimizzazione del processo di handover, viene presentato un modello analitico generale che caratterizza le prestazioni di un utente mobile che attraversa una rete eterogenea, in funzione della sua mobilità, del profilo di potenza ricevuta dalle celle vicine, del carico di traffico delle diverse celle, del tempo di trigger e del margine di isteresi. A differenza di molte soluzioni presenti in letteratura che mostrano solamente metodi euristici di ottimizzazione, in questa tesi viene sviluppata un'analisi Markoviana rigorosa per modellare il processo di handover considerando i possibili stati di un utente mobile lungo la sua traiettoria. Viene proposto poi un algoritmo ottimo basato sulle informazioni di contesto. Il modello matematico è validato tramite simulazioni in diversi scenari di trasmissione per confrontare le prestazioni della strategia presentata con le tecniche di handover convenzionali. Infine, viene investigato uno schema totalmente generale per calcolare le prestazioni massime di qualsiasi algoritmo di handover proposto in letteratura. La soluzione presentata è utile non solo per determinare il margine di miglioramento degli schemi esistenti di handover, ma anche perchè fornisce un'analisi comparativa tra le prestazioni ottenute dai vari algoritmi.

Per quanto riguarda l'ottimizzazione delle strategie di caching, viene studiato un modello di sistema estremamente innovativo perchè prevede la capacità di memorizzazione dei contenuti, ovvero il caching, sia nelle stazioni base delle celle sia negli utenti mobili. Di conseguenza, quando un qualsiasi utente richiede un contenuto, quest'ultimo può essere recuperato in diversi modi, attraverso una comunicazione device-to-device instaurata con

un altro utente, oppure attraverso il canale cellulare in downlink da una cella. Gli utenti sono perciò incoraggiati a condividere tra di loro i contenuti delle loro cache, ma possono anche stabilire una connessione tradizionale con la cella vicina se la ricerca del contenuto tra nodi paritari dovesse fallire. All'interno della rete eterogenea considerata, viene calcolato in forma chiusa il costo medio di sistema in funzione del costo associato alle singole operazioni di recupero del contenuto, del profilo di mobilità degli utenti e della distribuzione dell'interesse per i diversi contenuti. Viene implementata la minimizzazione del costo di sistema e tramite simulazione viene rilevato un significativo guadagno rispetto alle politiche di caching statiche, ovvero non dipendenti dal contesto. Nella maggior parte dei lavori proposti in letteratura si assume che gli utenti siano statici, ovvero privi di mobilità e quindi connessi alla medesima stazione base, e dotati dello stesso interesse per i contenuti, e quindi con la stessa probabilità di richiedere lo stesso insieme di contenuti. In questa tesi invece le assunzioni sono più realistiche, poichè si prevede che ciascun utente abbia la propria preferenza per un sottoinsieme di file.

List of Acronyms

3GPP 3rd Generation Partnership Project

BS Base Station

CAHP Context-Aware Handover Policy

CIO Cell Individual Offset

CSI Channel State Information

D2D Device-to-Device

GOP Group of Picture

HetNet Heterogeneous Network

HO Handover

ICN Information Centric Network

ISP Internet Service Provider

MC Markov Chain

QoS Quality-of-Service

RLF Radio Link Failure

RSRP Reference Signal Received Power

SINR Signal to Interference and Noise Ratio

SIR Signal to Interference Ratio

SNR Signal-to-Noise Ratio

SON Self Organizing Network

SSIM Structural Similarity (index)

SVM Support Vector Machine

TTT Time To Trigger

UE User Equipment

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The goal of this thesis is to design implementable algorithms for the resource management and optimization within Heterogeneous Networks (HetNets) [1]. These are networks consisting of various wireless access technologies, each of them having different capabilities, characteristics, constraints, and operating functionalities [2]. Specifically, micro, pico, and femto cells, as well as relay stations, can coexist in the same geographical area underlaying the macro cellular system, and their respective Base Stations (BSs) can potentially share the same spectrum. Since the deployment of small cells requires relatively low network overhead, HetNets can reduce the energy consumption of the future wireless networks.

Moreover, HetNets entail a significant paradigm shift, transitioning from a traditional and centralized macro cell approach to a more autonomous, distributed, and intelligent infrastructure. Small cells not only decrease the distance between BSs and end users, but also alleviate macro BSs by users' offloading, thus improving the indoor coverage, the cell-edge user performance, and the capacity of the overall network.

It is clear that in this kind of scenario, new network optimization challenges arise. Interference management, e.g., through power control and cell association schemes, is one of the major issues due to the unplanned deployment of small cells and their unpredictable working times [3]. Other examples of technical problems include cell selection and handover procedures that are essential to provide a seamless service when users move in or out of the cell coverage. Furthermore, efficient handovers are essential for traffic load balancing, by shifting users at the border of overlapping cells, from the more congested cells to the less congested ones. Backhaul network design is also a delicate task due to the complex topol-

ogy of the various types of coexisting cells. In fact, some cells may have dedicated interfaces to the core network, some others may form a cluster to aggregate and forward the traffic to the core, and others may rely on relays as an alternative interface [2].

From an information delivery perspective, HetNets enable also media content dissemination, as in Content Distribution Networks. According to this perspective, HetNets exploit geographically distributed services that transparently shift content from the origin servers to an optimized network of edge servers, or caches, located closer to the user who is requesting the content. The use of caches in fact improves the network performance, by minimizing latency and jitter, improving content accessibility, and balancing the server loads. Clearly, other challenges come up with this rationale. Some of the most important questions are related to where to place the edge servers, which content to outsource, which practice to use for the content replacement, and which route is the most effective to deliver the content from the appropriate server to the client requesting for it. The latter is also known as the path selection problem [4]. Since the storage capacity of mobile devices is typically limited, identifying which content a user should store in its cache, denoted as cache replacement problem [5], is one of the most crucial issues. Moreover, HetNets allow to exploit also the storage capacity of small BSs and relays, thus potentially improving the hit ratio and the system Quality of Service.

There are two main categories of routing and caching protocols, namely reactive and proactive protocols. On the one hand, reactive policies update the routing information and the stored content on-demand, only when contents are requested and routes need to be created or adjusted. On the other hand, proactive strategies periodically update the retrieved content and the routing information.

Among the above issues we have chosen the mobility management through the handover optimization and the proactive content replacement strategy as the two main investigations of this thesis. We introduce in Sec. 1.1 our motivation and contribution related to mobility management and handover optimization, while we present in Sec. 1.2 our main results on the proactive caching paradigm. The comprehensive analyses of the two topics are elaborated in Chapters 2 and 3 of this thesis, respectively.

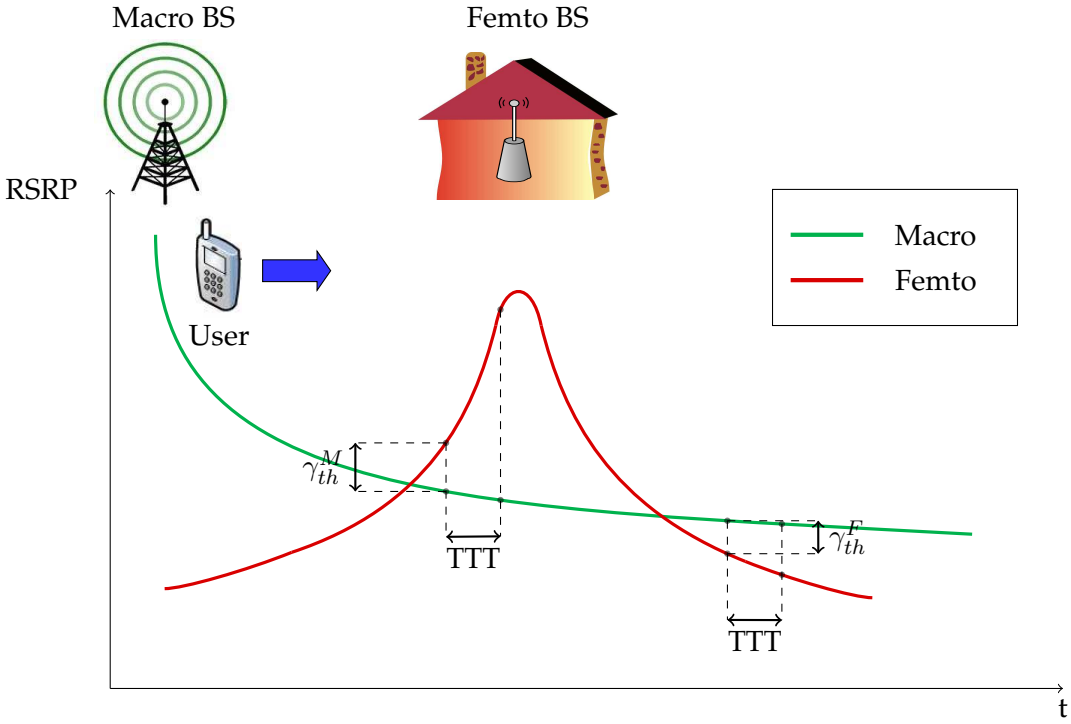

Figure 1.1. *Example of the decay of the power profile from the M-BS and F-BS as the UE moves away from the M-BS and towards the F-BS.*

1.1 Mobility Management through the Handover Process within HetNets

Global mobile data traffic is expected to increase exponentially in the next years, reaching 15.9 exabytes per month by 2018 [6]. One of the most promising approaches to face this challenge is the so-called HetNet paradigm, which basically consists in enriching the current cellular network with a number of smaller and simpler BSs, having widely varying transmit powers, coverage areas, carrier frequencies, types of backhaul connections and communication protocols. The deployment of pico and/or femto BSs *within* the macrocell, indeed, can provide higher connection speed and better coverage to the mobile users located at the border of the macrocell or in regions with high traffic demand.

While increasing the efficiency of the cellular networks, HetNets also raise several technical challenges related to user management [1]. An important aspect is related to the *handover* (HO) process of mobile users that, differently from classical cellular networks, have to deal with cells of widely varying coverage areas. In general, the HO process, standardized

by the 3rd Generation Partnership Project (3GPP) [7], is triggered by the User Equipment (UE), which periodically measures the Reference Signal Received Power (RSRP) from the surrounding cells. When the difference between the RSRP of a neighboring cell and that of the serving cell is higher than a fixed HO hysteresis value, γ_{th}, (event A3 in [8]), the HO process starts, as exemplified in Fig. 1.1. If this condition holds for a period of time equal to the *Time-To-Trigger* (TTT) parameter, the HO is finalized and the UE connects to the BS with the strongest RSRP.

The static setting of the HO hysteresis and TTT values adopted in traditional scenarios with only macrocells is no longer effective for HetNet systems, because of the large variety in cell characteristics [9,10]. With large values of TTT and of the hysteresis margin, the UE will likely experience a severe degradation of the RSRP during the TTT period when crossing a small cell, a problem that is generally referred to as *HO Failure*. On the other hand, short TTT and low hysteresis margin may cause *HO Ping-Pong*, i.e., frequent HOs to/from the M-BS, which yields performance losses due to signaling overhead and handover times. Reducing HO failure and ping-pong rates are clearly conflicting objectives, and the HO policy needs to trade off the two aspects [11].

Another challenge of HetNet management is the so called *Load Balancing*, which consists in mitigating congestion in cellular networks by offloading users from overloaded cells to lightly loaded neighboring cells. This problem has been mostly addressed in homogeneous networks, with only macrocells. Load Balancing in HetNets is more involved due to the disparities in cell sizes and transmit powers. In order to achieve the desired efficiency from the deployment of small cells, hence, the handover decision needs also to be load-aware. Indeed, by properly adapting the hysteresis margin, mobile users may be encouraged to switch to small BSs that are lightly loaded to get higher data rates. As a consequence, macrocells will also have the possibility to better serve their remaining users.

In this thesis, we make a step forward towards the design of context-aware HO policies by considering a basic but representative HetNet scenario where the mobile UE is crossing a small cell deployed at the edge of the macro cell coverage area. We first compute the closed form expression of the performance experienced by the UE (Sec. 2.3.1). This has been derived under a simple pathloss channel model that does not include multipath and shadowing effects. In this simplified scenario, we derive a preliminary version of a Context-

Aware Handover Policy (CAHP) showing the importance of binding the HO procedure to the context parameters (Sec. 2.3.2). We then consider a new system model that accounts for Rayleigh fading, and present a theoretical approach that describes the evolution of the UE state along its trajectory through a Markov chain (Sec. 2.4). We determine the expression of the average UE performance as a function of the HO parameters and other context parameters, such as the UE speed, the power profiles of the macro/pico/femto BSs, the cell load factors, and the channel profile (Sec. 2.5-Sec. 2.6). The mathematical framework we developed can accommodate different performance metrics, such as the HO failure rate, the ping-pong rate, or the average Shannon capacity, which is the one actually considered in this work. The model is then used to design the new version of CAHP (Sec. 2.7-Sec. 2.8) that selects the HO parameters to maximize the performance metric with respect to the UE environment and channel conditions.

Finally, we investigate a benchmark to compare our proposed HO algorithm with the upper bound case (Sec. 2.9). We propose a mathematical framework for the performance analysis of HO algorithms which is general enough to accommodate different context parameters and performance indices. Furthermore, we propose a Markov model that makes it possible to derive the *optimal HO performance* under the assumption that the channel conditions experienced by the mobile UE along its trajectory are known in advance. The performance of such a HO algorithm, theoretically achievable with suitable channel state information, can hence be used as a benchmark to compare the performance of different practical algorithms and assess their remaining room for improvement [12].

Chapter 2[1] presents our results that can be found in [13–16].

1.2 Proactive Content Replacement Policies within HetNets

The data traffic from wireless mobile devices is increasing worldwide, and will exceed the wired traffic by 2019, reaching 66% of the total Internet traffic, according to a recent study on mobile data usage by Cisco [17]. This rapid increase in mobile data activity raises the challenge of developing new technologies that can efficiently support this huge traffic demand.

[1] A preliminary version of the results presented in this Chapter has been done in collaboration with Francesco Guidolin, Ph.D. student from the University of Padova.

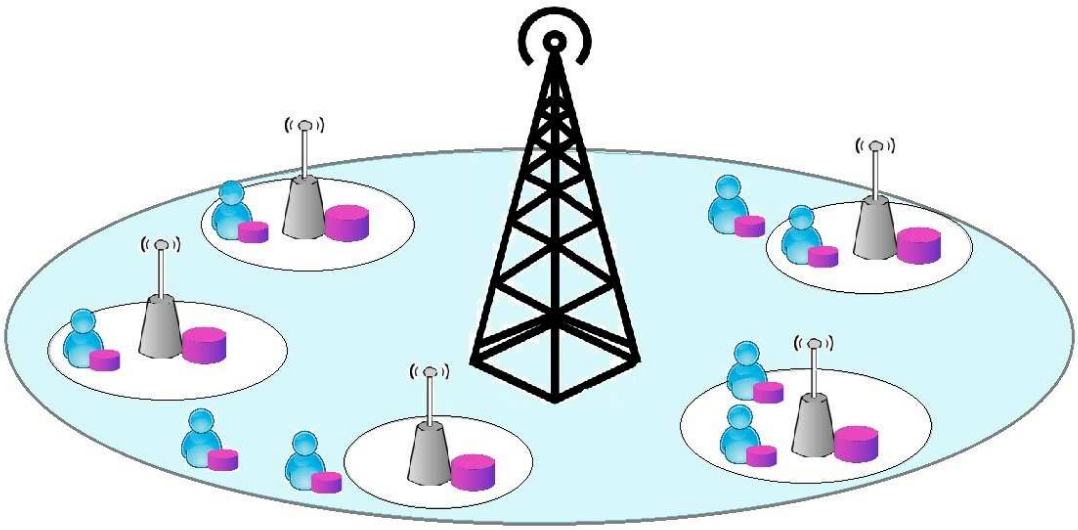

Figure 1.2. *Scenario for cooperative caching in a heterogeneous network, where small cells are deployed within a macro cell. Both mobile users and small BSs are provided with caches and may deliver the requested content to the neighboring users.*

An interesting trend is developing in the framework of heterogeneous networks (HetNets), i.e., networks with different access technologies. In a HetNet, the deployment of small cells base stations (BSs) is a cost-effective solution to offload data traffic from the macro cell network [18]. Currently, the fraction of smartphone traffic that is offloaded is about 46%, and will increase to 54% by 2019 [17]. At the same time, the weak backhaul links of the newly deployed small BSs become the bottleneck for wireless transmissions, and can reduce the advantages of the HetNet architecture. This is critical especially during peak hours, when links are congested and network resources are scarce [19]. To overcome this issue, a promising approach is to exploit the storage capacity of each small BS [20], which can serve with local data the users' requests, thereby decreasing the traffic through its backhaul link or the macro BS.

In this thesis we push this trend even further by taking advantage of the storage capabilities of other users in the system, assuming that all users are willing to cooperate. An example of the used reference scenario is depicted in Fig. 1.2. A file request can hence also be served by another user through a device-to-device (D2D) communication, without requiring a connection to the small or macro cell. In fact, the previously downloaded files

can be stored in the users' cache and shared with other users in an opportunistic and cost effective fashion. This approach allows for high spectral reuse, due to the short distance communication, and makes serving a file request faster and cheaper, since the user device has access to a large and collaborative virtual cache of files.

According to standard reactive caching policies, contents are retrieved to satisfy each user requests after they are initiated. A new caching paradigm introduced recently in the literature follows the so called *proactive* principle. The idea is to predict the possible requested files in the near future and pre-fetch them at local caches before they are actually requested. Motivated by the fact that traffic activity often exhibits a predictable pattern [21], we consider in our scenario a proactive caching policy as a promising possibility to outperform the reactive paradigm.

We hence develop a joint caching policy to exploit the storage capacity both at the terminal nodes and at the small cells. In the literature, a rigorous study of this type of systems is still lacking. A framework of this kind was recently introduced by [22], with a general description of a similar system design, but without a mathematical model. In this work instead we investigate the aforementioned system model, provide a closed form expression for the average system cost and derive a robust optimization framework for caching.

Defining an efficient caching policy is not trivial and often leads to the formulation of NP-hard problems as in [20, 23, 24]. An intuitive policy works by storing the most popular files at the caches [25], but in our scenario this policy is suboptimal since it does not jointly optimize the collaborative virtual caching among users and the small and macro BSs caches. Moreover, since the caching space is limited, a dynamic policy that is aware of the current users' preferences needs to be designed. Indeed, in contrast to many studies in the literature that assume that all users have the same traffic activity, i.e., they are interested in the same globally popular files, we also assume different classes of user interests, each of which follows a specific content popularity distribution. In fact, users may not value contents in the same way and may not be interested in the same set of contents. Moreover, they might follow diverse mobility patterns, and have different opportunities to meet other users and hence to retrieve contents from them.

Our main contributions on proactive caching management are summarized as follows. We first compute the probabilities that a requested file is available through a D2D com-

munication, a downlink with the closest small cell, and a downlink with the macro cell, respectively (Sec. 3.3). Secondly, we find the average system cost as a function of the user mobility pattern, the distribution of file interests, and the system variables that characterize which content each user and BS should store. Based on the above problem formulation, we find the optimal content allocation by minimizing the average system cost, thus developing our optimal caching policy. Due to the complexity of the optimization problem we formulate a suboptimal caching policy whose performance is proven to be sufficiently close to the optimal one (Sec. 3.4). Finally, we analyze through a vast simulation campaign how the context conditions, e.g., user mobility level, skewness of content popularity, and user interest profile, influence the performance of the proposed strategies (Sec. 3.5).

Chapter 3 presents our results that can be found in [26, 27].

1.3 Content-Aware Video Resource Allocation

As a final contribution of this thesis we report in Appendix C a preliminary analysis that has been developed as part of the Ph.D. program, in collaboration with Giulio Ministeri from the University of Padova.

The work focuses on a content-based optimization for video resource allocation using Bayesian Networks. We aim at developing a learning system that can automatically predict the quality-rate characteristic of an unknown video from its frame sizes. Since the video quality varies differently according to the dynamics of the scenes, a dynamicity classifier and predictor can be useful for an optimal content-aware resource allocation of the video. Hence, the first objective of this work, that is the one considered in Appendix C, is the inference of video dynamicity by analyzing the sizes of a certain number of frames. This information can then be used in a subsequent part of this work for the design of a video resource allocation scheme.

Chapter 2

Context-Aware Handover Policies in Heterogeneous Networks

This chapter is organized as follows. Sec. 2.1 provides an overview of prior work on the handover policies in the literature. Sec. 2.2.1 introduces the two channel propagation models used for the analysis. The former is a simple pathloss model while the latter includes also the fading component. Sec. 2.2.2 describes the handover mechanism, while Sec. 2.2.3 derives the user performance metric. Under the assumption of the pathloss only propagation model, in Sec. 2.3 we compute the closed form expression of the average performance, derive a preliminary version of our Context-Aware Handover Policy (CAHP) obtained from such a model, and validate the policy through simulations. Sec. 2.4 presents the analysis of the handover process under the more general pathloss plus fading channel model. In Sec. 2.5 we propose a Markov-based framework to model the user state during the handover process by means of a discrete time Markov chain, while in Sec. 2.6 we extend this model considering also the cell loads. Sec. 2.7 formulates the general version of CAHP, while Sec. 2.8 provides some evaluation results for different scenarios, in comparison also with other standard strategies. In Sec. 2.9 we evaluate a general framework to derive the upper bound performance of handover in HetNets. Finally, we discuss in Sec. 2.10 how the model described in Sec. 2.5 can be extended to a multicell scenario.

2.1 Prior Work

Recent surveys on self-organizing networks (SON) [28] and on mobility management in HetNets [12] clearly show that a proper configuration of the system parameters is both crucial for the overall throughput and also challenging due to the heterogeneity of the network. Some works in the literature focus on the theoretical characterization of key handover performance metrics. The authors in [29] express the relation between HO failure and ping-pong rates as a function of TTT, hysteresis margin, and user velocity. Similarly, in [30] the HO failure probability is derived as a function of the sampling period used by the user to collect the measurements from the neighboring cells, i.e., the Layer 3 filtering period. In both works however fast fading and shadowing statistics are neglected in the propagation model. In [31], instead, a closed-form expression of the HO failure rate is provided, taking into account also channel fading. The most severe limitation of the works in [29–31] is the assumption that small coverage areas are modeled as perfect circles that, while allowing their analytical tractability, is quite unrealistic. A study of more general user trajectories is presented in [32], where the authors propose a realistic user mobility model, and present analytic expressions for the HO rate, i.e., the expected number of HOs per unit time, and the cell sojourn time, i.e., the expected duration that the user stays within a particular serving cell.

Several solutions in the literature consider to adapt some HO parameters to the UE mobility conditions. In [33], for instance, the authors propose an algorithm that, while keeping the TTT and hysteresis margin constant, adaptively modifies the Cell Individual Offset (CIO) parameter, which is a margin to be added to the RSRP for load management purposes. The authors show that a UE can detect changes in its mobility pattern by monitoring the changes of the type of HO failure events (e.g., too early/late HO events, HO failures, or HO to the wrong cell) and, hence, can adjust the specified CIO parameter to minimize both the HO failure and the ping-pong rates.

In [34] an extensive simulation campaign is conducted in SONs to compute the Radio Link Failure (RLF)[1] rate for different UE speeds and types of handover, i.e., macro-to-macro and macro-to-pico handover. The proposed policy selects the TTT parameter that guaran-

[1] According to the standard [7], a RLF is declared when the user SINR remains below a certain threshold Q_{out} for a specified amount of time (usually 1 s).

tees that the RLF rate is below a certain threshold. Reference [35] analyzes the Cell Range Expansion (CRE) technique that consists in enlarging the small cell coverage in order to balance the users load. The authors simulate the effect of both CRE bias and hysteresis margin on the HO failure and ping-pong rates, while fixing the TTT parameter.

A different approach is presented in [36] where the HO decision is based on a mobility prediction algorithm that estimates the residence time of the UE in the possible target cell. The proposed policy allows the UE to switch to the target cell only when the estimated residence time is above a certain threshold. A similar procedure is considered in [37] where a mobility state estimation algorithm groups UEs into three speed classes and assigns a fixed TTT value to each of them, such that high speed UEs avoid the HO to pico cells, while lower speed UEs perform HO in order to minimize their RLF rate.

In these works, however, all users are assumed to have full access to the entire cell resources, irrespective of the current traffic load of each cell, which is unrealistic. The load balancing problem has been studied in [38], where the authors analyze the impact of the CRE parameter on the system capacity through the Cumulative Distribution Function (CDF) of the Signal to Interference plus Noise Ratio (SINR). The CRE parameter is adjusted to control the number of off-loaded users and, hence, to guarantee that the overall capacity is maximized. However, [38] assumes static users and does not take into account the handover that arises with mobile users. The algorithm in [39], instead, exploits the user mobility state and, by properly changing the users CIO parameter, reduces the congestion of overloaded cells, but without optimizing TTT and the hysteresis margin. The procedure described in [40] studies the impact of both the hysteresis margins referred to HOs to macro and small cells, assumed different in general, on the HO signaling overhead while guaranteeing the load balancing condition among users. The authors of [41], instead, propose a joint algorithm that, on the one hand, tunes TTT and the hysteresis parameters to optimize the handover performance metric (defined as a weighted sum of RLF, ping pong and handover failure) and, on the other hand, adapts the handover margin to achieve a load balancing condition.

Although these solutions improve the efficiency of HO in HetNets with respect to the standard static setting of the HO parameters, to the best of our knowledge a mathematical model that describes the HO performance as a function of the scenario parameters, such as the pathloss coefficients, the UE speed, and the cell load factors, is still lacking. In [14] we

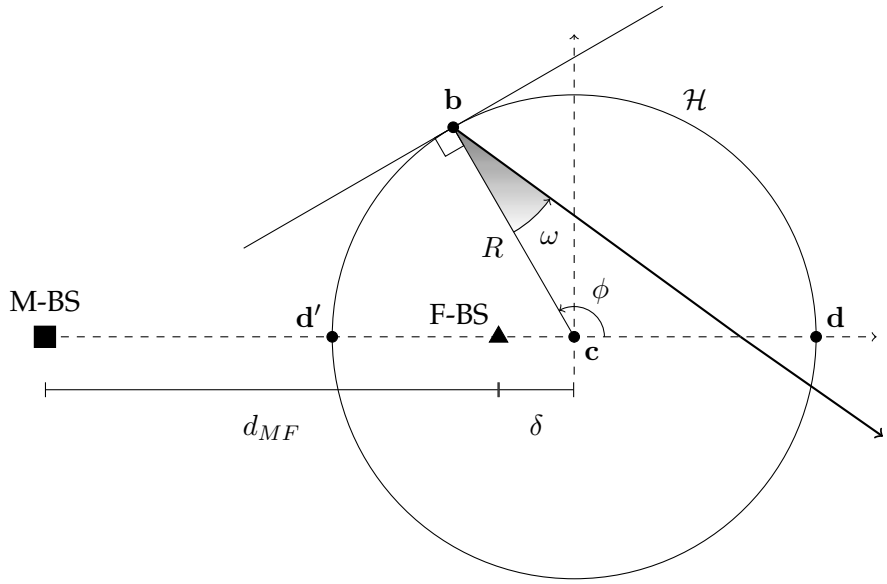

Figure 2.1. *Reference scenario: macrocell BS – M-BS (■), femtocell BS – F-BS (▲), and HO line $\mathcal{H}$ approximated as a circle of radius R and center* **c**. *Linear trajectory followed by a UE when entering the femtocell at point* **b** *with incidence angle ω.*

addressed this gap by proposing an approximate analytical expression for the mobile UEs performance, which is then used to define a TTT selection policy that maximizes the average Shannon capacity perceived by the UE along its trajectory. However, fading effects and load balancing conditions were not considered.

A similar work with respect to the one described in this thesis has been proposed in [42], where the authors develop a mathematical model for the HO procedure and derive a closed-form expression of the UE outage probability. Their policy selects the TTT and margin parameters in order to minimize the specific metric of handover failure rate. However, they do not consider the problem of load balancing among cells and, moreover, make the assumption that the UE trajectory with respect to the position of the BSs is known to the UE. Our work, instead, proposes a more general model, and defines a context-aware HO strategy based on the more realistic assumption that the UE's trajectory with respect to the location of the BSs is unknown and that the cells are loaded.

2.2 System Model

For the sake of simplicity, we focus on a basic scenario consisting of a macro BS (M-BS) and a femto BS (F-BS) placed at distance d_{MF}, and using the same frequency band. Despite its simplicity, this model still presents the fundamental issues related to HO in HetNets and, hence, is representative of the targeted scenario. In any case, the approach we propose in this manuscript can be generalized to more complex scenarios with multiple overlapping femtocells, though at the cost of a more involved notation and argumentation, as discussed in Sec. 2.10.

For convenience, we define the UE's trajectory with respect to a reference circle $\mathcal{H}$ of radius R centered at the F-BS. We adopt the model proposed in [29] that approximates $\mathcal{H}$ as a circumference of radius R centered in a point $\mathbf{c}$ at distance δ from the F-BS in the opposite direction with respect to M-BS, as shown in Fig. 2.1. We assume that the UE moves at constant speed v, following a straight trajectory. With reference to the polar coordinate system depicted in Fig. 2.1, the trajectory is then uniquely identified by the angular coordinate ϕ of point $\mathbf{b}$ where the UE crosses the border $\mathcal{H}$, and by the incidence angle ω formed by the trajectory with respect to the radius passing through $\mathbf{b}$. As done in [29], we assume that the UE can enter the femtocell from any point and with any angle, so that the parameters ϕ and ω are modeled as independent random variables with uniform distribution in the intervals $[0, 2\pi]$ and $[-\pi/2, \pi/2]$, respectively.

2.2.1 Propagation model

At time t, a mobile UE at position $\mathbf{a}$ measures an RSRP $\Gamma_M(\mathbf{a}, t)$ from the M-BS, and $\Gamma_F(\mathbf{a}, t)$ from the F-BS. We initially assume a simple path-loss channel model that does not include multipath and shadowing effects.

The average power received in position $\mathbf{a}$ from h-BS can then be expressed as [43]

$$\Gamma_h(\mathbf{a}) = \Gamma_h^{tx} \left(\frac{d_h(\mathbf{a})}{d_{0h}} \right)^{\eta_h} \quad h \in \{M, F\}\,, \tag{2.1}$$

where Γ_h^{tx} is the transmit power of h-BS, η_h the path loss exponent, d_{0h} the reference distance for the far field model to apply, and $d_h(\mathbf{a})$ the distance of point $\mathbf{a}$ to h-BS, with $h \in \{M, F\}$. In the following we will assume $\eta_F \leq \eta_M$, as usual in the literature [44].

Since the considered scenario is interference-limited, we can neglect the noise term and approximate the SINR experienced by a UE in position $\mathbf{a}$ as

$$\gamma_M(\mathbf{a}) = \frac{\Gamma_M(\mathbf{a})}{\Gamma_F(\mathbf{a})}, \quad \gamma_F(\mathbf{a}) = \frac{\Gamma_F(\mathbf{a})}{\Gamma_M(\mathbf{a})}, \tag{2.2}$$

when it is connected to the M-BS and the F-BS, respectively. The analysis in Sec. 2.3 adopts (2.1) as the propagation model, while from Sec. 2.4 on we consider a more general model with a path-loss plus fading propagation model [43]. According to the latter, the RSRP from the h-BS, with $h \in \{M, F\}$, is given by

$$\Gamma_h(\mathbf{a}, t) = \Gamma_h^{tx}\, g_h(\mathbf{a})\, \alpha_h(t), \tag{2.3}$$

where $g_h(\mathbf{a})$ is the pathloss gain, which depends only on the distance of point $\mathbf{a}$ from the h-BS, while $\alpha_h(t)$ is the fast-fading channel gain at time t. We assume that the fading is Rayleigh distributed, i.e., $\alpha_h(t)$ is an exponential random variable with unit mean and coherence time [45]

$$T_c = \sqrt{\frac{9}{16\pi}}\frac{1}{f_d} = \sqrt{\frac{9}{16\pi}}\frac{c}{v f_c}, \tag{2.4}$$

where f_d and f_c are the Doppler and the carrier frequencies, respectively, c is the speed of light, and v is the UE's speed. Due to fading, channel fluctuations can cause the HO process to be improperly triggered, thus generating the ping-pong effect. The duration of the channel outage is a well studied metric in the literature to model this phenomenon (e.g., see [46, 47]).

With (2.3) the SINR $\gamma_h(\mathbf{a}, t)$ experienced by a UE connected to the h-BS at time t and in position $\mathbf{a}$ is given by[2]

$$\gamma_h(\mathbf{a}, t) = \bar{\gamma}_h(\mathbf{a})\xi_h(t), \quad h \in \{M, F\}, \tag{2.5}$$

where

$$\bar{\gamma}_M(\mathbf{a}) = \frac{\Gamma_M^{tx}\, g_M(\mathbf{a})}{\Gamma_F^{tx}\, g_F(\mathbf{a})}, \quad \bar{\gamma}_F(\mathbf{a}) = \frac{\Gamma_F^{tx}\, g_F(\mathbf{a})}{\Gamma_M^{tx}\, g_M(\mathbf{a})}, \tag{2.6}$$

are the deterministic components of the SINR, while

$$\xi_M(t) = \frac{\alpha_M(t)}{\alpha_F(t)}, \quad \xi_F(t) = \frac{\alpha_F(t)}{\alpha_M(t)}, \tag{2.7}$$

account for the random variations due to fading.[3]

[2] The model can be extended to account for the interference from other cells, though for the sake of simplicity here we neglect other interference sources.

[3] In the simulations of Sec. 2.8, we relax the interference-limited assumption and take noise into consideration.

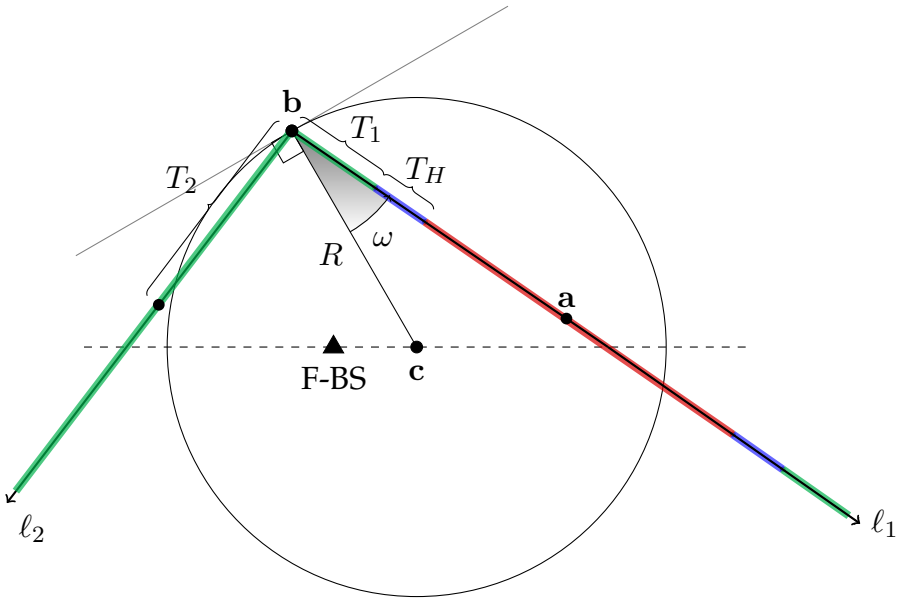

Figure 2.2. *Impact of the TTT timer on the user performance. Along ℓ_1 T_1 expires while the RSRP from the femtocell is higher than that from the macrocell and the HO is performed. On the contrary, along ℓ_2 T_2 expires after the UE exits the femtocell, when the RSRP from the macrocell is prevalent again, and the HO is avoided. Colors indicate the UE state: connected to the macrocell (green), connected to the femtocell (red), and switching from one to the other (blue).*

2.2.2 Handover performance model

The HO process is driven by the UE's instantaneous RSRP. If the difference between the RSRP of the serving and the target cell drops below the HO threshold γ_{th}, the TTT timer is initialized to a certain value T and the countdown starts. Whenever the RSRP difference returns above the HO threshold, however, the countdown is aborted and the HO procedure is interrupted. Conversely, if it remains below the threshold for the entire interval T, then the UE disconnects from the serving BS and connects to the new BS. This switching process takes a time T_H that accounts for the network procedures to connect the UE to the target BS. We depict in Fig. 2.2 two choices of T for which the HO is performed (trajectory ℓ_1) and avoided (trajectory ℓ_2). We remark here that the above condition on the RSRP difference can be translated to an equivalent condition on the SINR experienced by the UE where the power received from the target cell is the interference. Hence, we will use this latter notation in the following.

2.2.3 Mean Trajectory performance

For any given point $\mathbf{a}$, we can then define the connection state S of the UE to be M, F or H depending on whether the UE is connected to the **M**-BS, the **F**-BS or is temporarily disconnected because **H**anding over from one to the other.

Given an arbitrary straight path ℓ, we define the mean trajectory performance as

$$\mathcal{C}_\ell = \frac{1}{|\ell|} \int_\ell \sum_{S \in \{M,F,H\}} C_S(\mathbf{a}) \chi_{\mathbf{a}}(S) \, \mathrm{d}\mathbf{a} \,, \tag{2.8}$$

where $|\ell|$ is the trajectory's length, $\int_\ell$ is the line integral along the trajectory, $\chi_{\mathbf{a}}(S)$ is 1 if the UE's state at point $\mathbf{a}$ is S and zero otherwise, while $C_S(\mathbf{a})$ is the performance experienced by the UE at point $\mathbf{a}$ along the trajectory, given that it is in state $S \in \{M, F, H\}$. We remark here that $C_S(\mathbf{a})$ can be any arbitrarily chosen metric along the UE's trajectory.

C_ℓ strongly depends on the TTT value (see Fig. 2.2) since short TTT values increase the chance of HO, thus improving the SINR in the femtocell at the cost of the zero capacity penalty during the period T_H; on the other hand, large TTT values may let the UE cross the femtocell without switching to F-BS, thus suffering a lower SINR inside the femtocell, but avoiding the loss due to T_H.

Since the UE can follow any trajectory, we average the capacity along all the straight lines of length L that enter the femtocell with random incidence angle, thus obtaining[4]

$$C_L = \frac{2}{L\pi} \int_0^{\pi/2} \int_0^L \sum_{S \in \{M,F,H\}} C_S(\mathbf{a}(x,\omega)) \chi_{\mathbf{a}(x,\omega)}(S) \, \mathrm{d}x \, \mathrm{d}\omega \,, \tag{2.9}$$

with $\mathbf{a}(x, \omega)$ being the point at distance x from $\mathbf{b}$ along the trajectory with incidence angle ω.

In this thesis, we consider the average Shannon capacity experienced by the UE while crossing the femtocell as the performance metric $C_S(\mathbf{a})$, so that we define

$$\begin{aligned} &C_M(\mathbf{a}) = \log\left(1 + \gamma_M(\mathbf{a})\right) \,;\; C_F(\mathbf{a}) = \log\left(1 + \gamma_F(\mathbf{a})\right) \,; \\ &C_H(\mathbf{a}) = 0 \,; \end{aligned} \tag{2.10}$$

where $\gamma_M(\mathbf{a})$ and $\gamma_F(\mathbf{a})$ are given either in (2.2) or in (2.5), depending on the propagation model used. Note that we assign zero capacity during the actual switching from one BS

[4]For the symmetry of the problem, the entrance point $\mathbf{b}$ is irrelevant.

to the other one (state H) in order to account for the various costs of the handover process (energy, time, signaling, etc).

2.3 Handover performance under a pathloss propagation model

In Sec. 2.3.1 we express the average trajectory capacity (2.9) in a semi closed form, as a function of the TTT parameter and the UE speed. The used propagation model is given in (2.1). We assume for simplicity that $\gamma_{th} = 0$ dB.[5] Therefore, the HO starts whenever the mobile UE crosses the closed line $\mathcal{H}$ formed by the points $\mathbf{a}$ such that $\Gamma_M(\mathbf{a}) = \Gamma_F(\mathbf{a})$. Parameters R and δ can be found by setting

$$\Gamma_M(\mathbf{a}) = \Gamma_F(\mathbf{a}), \quad \text{for } \mathbf{a} \in \{\mathbf{d}', \mathbf{d}\}, \tag{2.11}$$

where points $\mathbf{d}'$ and $\mathbf{d}$ are shown in Fig. 2.1. Using (2.1) into (2.11), we get

$$\begin{cases} d_{0M}^{\eta_M} d_{0F}^{-\eta_F} 10^{\frac{\Delta\Gamma_{MF}}{10}} = \frac{(d_{MF}-(R-\delta))^{\eta_M}}{(R-\delta)^{\eta_F}} \\ d_{0M}^{\eta_M} d_{0F}^{-\eta_F} 10^{\frac{\Delta\Gamma_{MF}}{10}} = \frac{(d_{MF}+(R+\delta))^{\eta_M}}{(R+\delta)^{\eta_F}} \end{cases} \tag{2.12}$$

where $\Delta\Gamma_{MF} = \Gamma_M^{tx} - \Gamma_F^{tx}$. The solutions R and δ of (2.12) can be easily obtained with numerical methods.

We identify the *femtocell* as the area inside the circle $\mathcal{H}$, while the macrocell includes the femtocell and the surrounding area. When the UE connected to M-BS enters the femtocell, a TTT timer is initialized to the value T. If the UE exits the femtocell before the timer expires, the HO process is interrupted and the UE stays connected to M-BS. Conversely, if the timer expires while the UE is still in the circle, the HO is actually performed and the UE disconnects from the M-BS and connects to the F-BS in a time T_H. Similarly, when a UE connected to F-BS exits the femtocell, another HO process is started to connect back to the M-BS. We remark here that, without using the penalty time T_H, the optimal strategy would obviously consist in performing HO with zero TTT any time the SINR condition $\Gamma_M(\mathbf{a}) = \Gamma_F(\mathbf{a})$ is met. Finally, we assume that T is the same for both macro-to-femto and femto-to-macro HOs, though the analysis can be easily generalized to different Ts.

[5] Our model can be generalized to nonzero hysteresis margins, but at the cost of a more involved notation and analysis.

2.3.1 Closed form expression of the trajectory capacity

It is convenient to express (2.9) as

$$C_L = C_{L,int} + C_{L,ext}\,, \tag{2.13}$$

where

$$C_{L,int} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^{\pi/2} \int_0^{2R\cos\omega} C_S(\mathbf{a}(x,\omega))\chi_{\mathbf{a}(x,\omega)}(S)\mathrm{d}x\,\mathrm{d}\omega \tag{2.14}$$

is the contribution to the average capacity due to the part of the trajectory inside the femtocell, while

$$C_{L,ext} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^{\pi/2} \int_{2R\cos\omega}^{L} C_S(\mathbf{a}(x,\omega))\chi_{\mathbf{a}(x,\omega)}(S)\mathrm{d}x\,\mathrm{d}\omega \tag{2.15}$$

is the contribution of the part external to the femtocell. In the following, we work out each part separately.

2.3.1.1 Internal component

Let us now focus on (2.14). Under the simplified circular model for the femtocell, the SINR depends only on the distance a of point $\mathbf{a}(x,\omega)$ from the femtocell center. Given any circle of radius $a \leq R$ centered in $\mathbf{c}$, the trajectory can either cross it in two points, or not cross it at all (the tangent case is neglected having zero probability). In case of crossing, we denote by

$$\mathrm{d}_\pm = R\cos\omega \pm \sqrt{a^2 - R^2\sin^2\omega} \tag{2.16}$$

the length of the trajectory when it intersects the circle. In Fig. 2.3 we report the reference scenario to compute (2.14) where $\mathbf{d}_\pm$ are the two points where the UE has traveled $\mathrm{d}_\pm$ along its trajectory. Then, by changing variable x with $a = \sqrt{x^2 + R^2 - 2xR\cos\omega}$, (2.14) can be written as

$$C_{L,int} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^{\pi/2} \int_{R\sin\omega}^{R} C_S(a) \frac{\psi(a,S)}{\sqrt{1-(R/a)^2\sin^2\omega}}\,\mathrm{d}a\,\mathrm{d}\omega\,, \tag{2.17}$$

where $C_S(a)$ denotes the capacity at distance a from the femtocell center when the UE's state is S, while $\psi(a,S)$ counts the number of intersection points at which UE's state is S, i.e.,

$$\psi(a,S) = \chi_{\mathrm{d}_-}(S) + \chi_{\mathrm{d}_+}(S)\,, \tag{2.18}$$

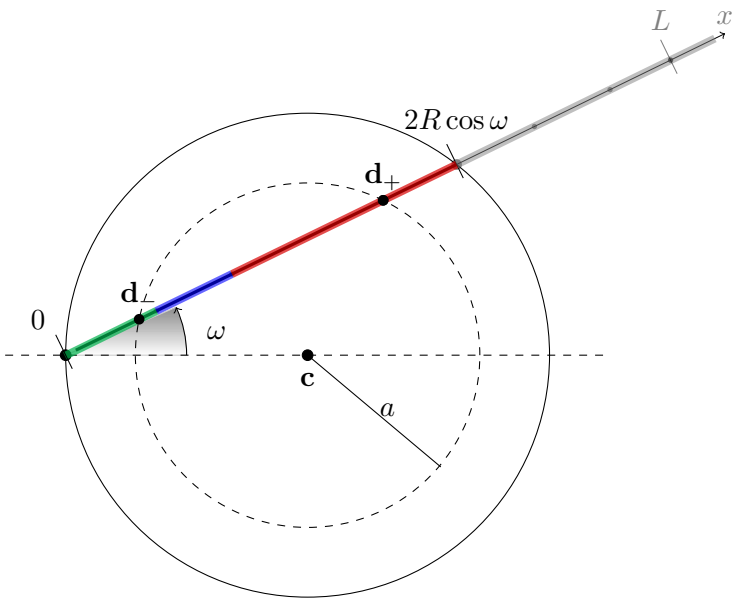

Figure 2.3. *Reference scenario to compute the internal component of the average capacity. Colors along the trajectory indicate the UE state: connected to the macrocell (green), connected to the femtocell (red), and switching from one to the other (blue).*

where $\chi_{\mathrm{d}_\pm}(S)$ is one if, after traveling a distance $\mathrm{d}_\pm$ along the trajectory, the UE's state is S, and zero otherwise. Changing the order of integration in (2.17) we get

$$C_{L,int} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^R C_S(a) \int_0^{\sin^{-1}(\frac{a}{R})} \frac{\psi(a,S)}{\sqrt{1-(R/a)^2\sin^2\omega}}\,\mathrm{d}\omega\,\mathrm{d}a\,. \tag{2.19}$$

Now, denoting by $y_T = v\,T$ and $y_H = v\,T_H$ the distance covered by the UE during the TTT time T and the handover time T_H, respectively, it is easy to realize that, for points within the femtocell, $\chi_d(M) = 1$ if $d < y_T$, $\chi_d(F) = 1$ if $d > y_T + y_H$, and $\chi_d(H) = 1$ otherwise. Therefore, for any given a, the inner integration interval in (2.19) can be split into subintervals $I_n(a,S) = [\alpha_n(a,S), \beta_n(a,S)]$, as specified in Table 2.1 for $S \in \{M,F\}$, where the function $\psi(a,S)$ is constant and equal to $n \in \{0,1,2\}$. The interval extremes are given

S	Coefficient $\psi(a,S)=n$ and conditions on y_T and y_H	$I_{n,S}(a)$ $[\alpha_n(a,S),\beta_n(a,S)]$
M	$n=1$	$[0,\omega_T(a)]$
	$n=0$ if $y_T\in[0,x_{tan}(a)]$	$[\omega_T(a),\omega_{max}(a)]$
	$n=2$ if $y_T>x_{tan}(a)$	$[\omega_T(a),\omega_{max}(a)]$
F	$n=1$	$[0,\omega_H(a)]$
	$n=0$ if $y_T+y_H\in[0,x_{tan}(a)]$	$[\omega_H(a),\omega_{max}(a)]$
	$n=2$ if $y_T+y_H>x_{tan}(a)$	$[\omega_H(a),\omega_{max}(a)]$

Table 2.1. *Integration intervals for the internal trajectory components. See (2.20)–(2.23) for the definition of the different functions.*

by

$$\omega_{max}(a)=\sin^{-1}(a/R)\,; \tag{2.20}$$

$$\omega_T(a)=\cos^{-1}\left\lfloor\frac{R^2+{y_T}^2-a^2}{2Ry_T}\right\rceil_0^1\,; \tag{2.21}$$

$$\omega_H(a)=\cos^{-1}\left\lfloor\frac{R^2+(y_T+y_H)^2-a^2}{2R(y_T+y_H)}\right\rceil_0^1\,; \tag{2.22}$$

$$x_{tan}(a)=\sqrt{R^2-a^2}\,; \tag{2.23}$$

with $\lfloor x\rceil_0^1=\min(1,\max(0,x))$. In practice, $\omega_{max}(a)$ is the incidence angle of the trajectory that is tangent to a circle of radius a centered on $\mathbf{c}$ and $x_{tan}(a)$ is the length at which this trajectory touches that circle, while $\omega_T(a)$ and $\omega_H(a)$ are the incidence angles for which the trajectory intersects the circle at distance y_T and y_T+y_H from the ingress point, respectively. Note that, for a given a, feasible values of $\psi(a,S)$ are either $\{0,1\}$ or $\{1,2\}$, depending on y_T and y_H. By using these intervals, after some algebraic steps reported in Appendix A.1, we can express the average capacity (2.19) as

$$C_{L,int}=\sum_{S\in\{M,F,H\}}\frac{2}{L\pi}\int_0^R C_S(a)\left[G\left(\beta_1(a,S),\beta_2(a,S),\frac{R}{a}\right)-F\left(\beta_2(a,S),\frac{R}{a}\right)\right]\mathrm{d}a\,, \tag{2.24}$$

where $G(\phi_1,\phi_2,k)=F(\phi_2,k)-F(\phi_1,k)$ with $F(\phi,k)$ being the *incomplete elliptic integral of the first kind*, which can be computed with standard methods.

S	Coefficient $\psi(a,S)=n$	$I_{n,S}(a)$ $[\alpha_n(a,S),\beta_n(a,S)]$
M	$n=0$	$[\omega_{min}(a),\tilde{\omega}_H(a)]$
	$n=1$	$[\max\{\tilde{\omega}_H(a),\omega_{min}(a)\},\pi/2]$
F	$n=0$	$[\min\{\tilde{\omega}_T(a),\omega^*\},\pi/2]$
	$n=1$	$[\omega_{min}(a),\min\{\tilde{\omega}_T(a),\omega^*\}]$

Table 2.2. *Integration intervals for the external trajectory components. See (2.26)–(2.27) and (2.29)–(2.30) for the definition of the different functions.*

2.3.1.2 External component

Following the same rationale used for the inner component, we can express (2.15) as

$$C_{L,ext}=\sum_{S\in\{M,F,H\}}\frac{2}{L\pi}\int_R^{\sqrt{R^2+L^2}}C_S(a)\int_{\omega_{min}(a)}^{\pi/2}\frac{\psi(a,S)}{\sqrt{1-(R/a)^2\sin^2\omega}}\,\mathrm{d}\omega\,\mathrm{d}a\,, \tag{2.25}$$

where $\psi(a,S)=\chi_{\mathrm{d}_+}(S)$ and

$$\omega_{min}(a)=\cos^{-1}\left\lfloor\frac{L^2+R^2-a^2}{2RL}\right\rceil_0^1. \tag{2.26}$$

Determining the values of $\psi(a,S)$ for points outside the femtocell is slightly more involved than in the previous case. We start observing that, if $\omega>\omega^*$, with

$$\omega^*=\cos^{-1}\left\lfloor\frac{y_T}{2R}\right\rceil_0^1, \tag{2.27}$$

then $S=M$ for any d, since the mobile leaves the femtocell before the time to handover T has elapsed. If $\omega\le\omega^*$, the state at distance d_+ is M only if the distance crossed outside the femtocell is larger than y_T+y_H.[6] Instead, we have $\chi_{\mathrm{d}_+}(F)=1$ if $\omega<\omega^*$ and the distance traveled outside the femtocell is less than y_T. In the other cases, we obviously have $\chi_{\mathrm{d}_+}(H)=1$.

As before, by splitting the inner integration interval in (2.25) into two subintervals $I_n(a,S)$ in which $\psi(a,S)$ is constant and equal to 0 or 1, we get

$$C_{L,ext}=\sum_{S\in\{M,F,H\}}\frac{2}{L\pi}\int_R^{\sqrt{R^2+L^2}}C_S(a)\,G\!\left(\alpha_1(a,S),\beta_1(a,S),\frac{R}{a}\right)\mathrm{d}a\,, \tag{2.28}$$

[6]Note that the HO can also start inside the femtocell and be completed after the UE has exited the femtocell. This case, which can be easily accounted for in the model, is quite cumbersome to be described and, hence, is not discussed further.

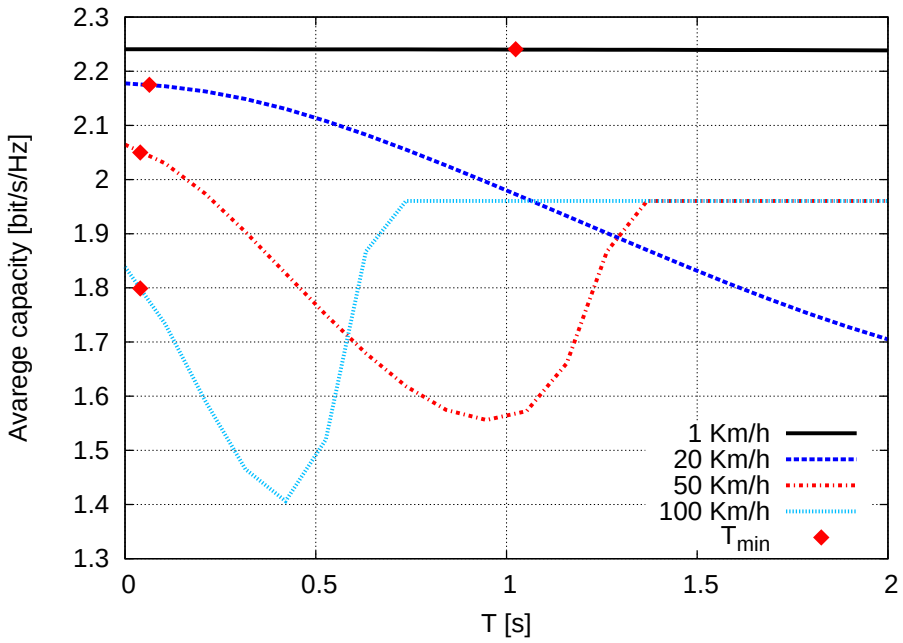

Figure 2.4. *Average capacity values vs TTT values, for various values of the mobile users speed.*

where the interval extremes in Table 2.2 are given by

$$\tilde{\omega}_T(a) = \cos^{-1}\left\lfloor \frac{a^2 - R^2 - {y_T}^2}{2Ry_T} \right\rceil_0^1 ; \tag{2.29}$$

$$\tilde{\omega}_H(a) = \cos^{-1}\left\lfloor \frac{a^2 - R^2 - (y_T + y_H)^2}{2R(y_T + y_H)} \right\rceil_0^1 . \tag{2.30}$$

2.3.2 Performance evaluation

Based on the model derived in Sec. 2.3.1, we analyze the impact of the choice of T on the selected performance index, namely, the average Shannon capacity, and we derive an optimal strategy, depending on the scenario parameters. Successively, we test our policy against different strategies by simulating the handover process in a more realistic scenario that also includes Rayleigh fading.

The numerical results have been obtained by setting $T_H = 200$ ms, and the transmit power of M-BS and F-BS to 46 dBm and 23 dBm, respectively [48].

Figure 2.4 shows the analytical capacity given by (2.13) as a function of T, for several values of the UE speed. We can see that, increasing the TTT, the average capacity first decreases and, then, increases again, till it reaches an asymptotic value that corresponds to the average capacity in case HO is never triggered. We remark that these results have been

obtained by considering the analytical model only and, hence, neglect the effect of fading, which will be discussed later. In this condition, the observed behavior reflects the balance between two opposite factors. On the one hand, a very short TTT favors the trajectories that have a significant internal component, i.e., that cut the cell close to its center and that experience a higher capacity if the UE switches to the F-BS as soon as possible. On the other hand, more peripheral trajectories suffer the loss due to the handover operations that is not compensated by the capacity gain obtained by connecting to the F-BS. When TTT increases, the capacity loss incurred by the inner trajectories dominates the gain of the peripheral trajectories, so that the net effect is a decrease of the average capacity. Above a certain TTT, this behavior changes and, for sufficiently large values of TTT, the capacity saturates since HO is never triggered.

We note that, for very low values of the UE speed, immediate handover is recommended because most trajectories will stay in the femtocell a time long enough to recover from the capacity loss incurred during the HO time T_H. The situation is the opposite for high speeds, from which it is better to set a very large TTT value to avoid HO.

This argument, however, neglects the effect of Rayleigh fading that, with very short TTT values, will likely result in severe ping-pong effect. Assuming signals are affected by independent Rayleigh fading processes, the ping-pong effect can be mitigated by setting T larger than a specific value, here denoted as T_{min}, which can be computed for each value of the UE speed using the results presented in [49]. We choose T_{min} to have a probability lower than 0.01 that the HO is improperly triggered by fading processes.

According to our mathematical analysis, then, the optimal handover strategy consists in either performing HO as soon as possible, i.e., setting $T = T_{min}$, or not performing HO at all, i.e., using $T = \infty$. This choice is bound to the context through the speed threshold v_{th}, below which HO is triggered with $T = T_{min}$, and above which HO is not triggered at all.

Fig. 2.5 shows the speed thresholds v_{th} for different combinations of η_F and η_M values. The speed threshold ranges from 1 to 180 Km/h, while the value 200 Km/h indicates that, in the considered scenario, the best strategy is to avoid HO. We note that the v_{th} trend depends on the specific pathloss exponents and not only on their ratio. Moreover, since the cell coverage is determined by the η_h values, we can notice that v_{th} is directly proportional to the cell size. Then, to optimize the handover procedure it is important to know the mobility

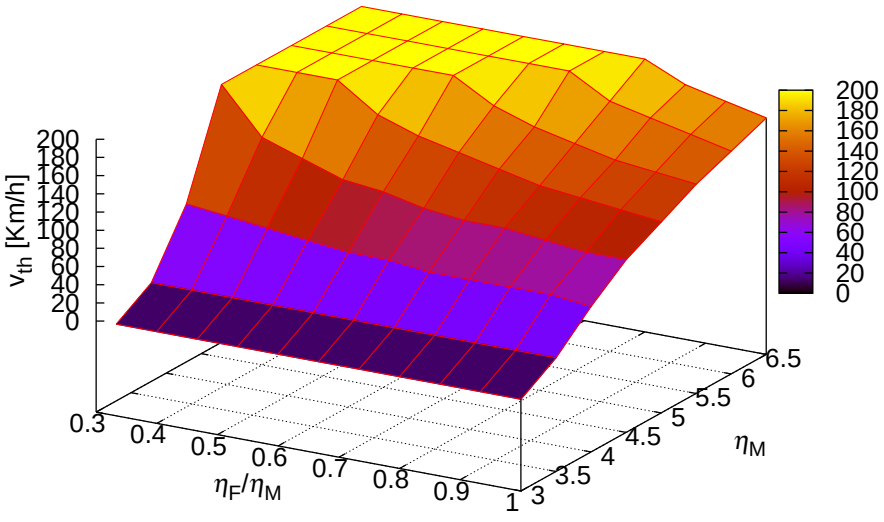

Figure 2.5. *v_{th} for different pathloss ratios.*

characteristics of the users and the channel parameters.

Finally, we compare via Monte-Carlo simulations the performance obtained with three different policies, i.e., [CAHP]: the context-aware policy that performs handover only when $v < v_{th}$ and, in that case, uses $T = T_{min}$; [FIX]: a policy with TTT fixed to $T = 100$ ms for every speed; [TMIN]: a minimum TTT policy, where $T = T_{min}$ for each speed value. The M-BS and F-BS are placed at a distance $d_{MF} = 500$ m. The signals received from M-BS and F-BS are generated according to two independent path-loss plus Rayleigh fading channel models, with path-loss exponents $\eta_M = 4$ and $\eta_F = 2$, respectively, and coherence time depending on the UE speed. For every Monte-Carlo run, we compute the average capacity experienced by a user that crosses the femtocell coverage area with a linear trajectory of length L equal to twice the macrocell radius.

Fig. 2.6 shows the average capacity obtained using the three different policies. Note that, at low speeds, the FIX strategy suffers from the ping-pong effect that determines a strong performance degradation. Conversely, TMIN and CAHP perform much better by allowing HO after the minimum TTT required to limit the ping-pong effect. For larger speed values, CAHP gains over TMIN because it skips HO, avoiding the loss due to the two T_H in a short time interval. In this case, also the FIX policy with $T = 100$ ms achieves the best

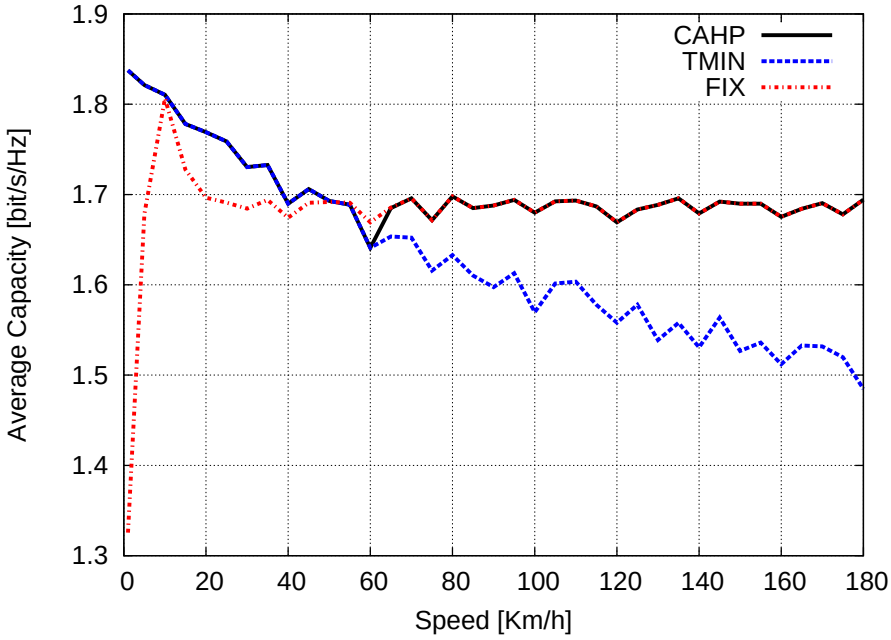

Figure 2.6. *Average capacity obtained with different approaches.*

performance, since the TTT is long enough to avoid handover for the considered scenarios, though performance may drop for other scenarios. The capacity fluctuations are due to the fast fading effect.

2.4 Handover performance under a pathloss plus fading propagation model

From this section on, we consider a general propagation model that includes fading, where the RSRP is given in (2.3). In this case, the computation of the average trajectory capacity (2.9) (reported here for convenience)

$$C_L = \frac{2}{L\pi} \int_0^{\pi/2} \int_0^L \sum_{S \in \{M,F,H\}} C_S(\mathbf{a}(x,\omega)) \chi_{\mathbf{a}(x,\omega)}(S) \, \mathrm{d}x \, \mathrm{d}\omega \ ,$$

is more involved due to the fading effect. In particular, the term $\chi_{\mathbf{a}(x,\omega)}(S)$ is random, depending on the evolution of the SINR in the previous time interval of length T. Taking the expectation of (2.9) with respect to the random variables $\xi_h(t)$, $h \in \{M, F\}$, defined in (2.7),

we hence get

$$\bar{\mathcal{C}}_L = \frac{2}{L\pi} \int_0^{\pi/2} \int_0^{L} \sum_{S \in \{M,F,H\}} \bar{C}_S(\mathbf{a}(x,\omega)) \mathrm{P}_S[\mathbf{a}(x,\omega)] \mathrm{d}x \, \mathrm{d}\omega \,, \tag{2.31}$$

where $\bar{C}_S(\mathbf{a}(x,\omega))$ is the average performance at point $\mathbf{a}(x,\omega)$, given that the UE's state at point $\mathbf{a}(x,\omega)$ is S, whose probability is

$$\mathrm{P}_S[\mathbf{a}(x,\omega)] = \mathbb{E}\left[\chi_{\mathbf{a}(x,\omega)}(S)\right] \,. \tag{2.32}$$

Since in our analysis $C_S(\mathbf{a})$ is the Shannon capacity experienced by the UE in state S, we compute $\bar{C}_S(\mathbf{a})$ as the *average* Shannon capacity with respect to the fading component. Hence, for $S \in \{M, F\}$ we define

$$\begin{aligned} \bar{C}_S(\mathbf{a}) &= \mathbb{E}\left[\log_2\left(1 + \gamma_S(\mathbf{a}, t)\right)\right] \\ &= \log_2\left(\bar{\gamma}_S(\mathbf{a})\right) \frac{\bar{\gamma}_S(\mathbf{a})}{\bar{\gamma}_S(\mathbf{a}) - 1} \,, \end{aligned} \tag{2.33}$$

where the expression in the last row is derived in Appendix A.2. As before, during T_H we assume

$$\bar{C}_H(\mathbf{a}) = 0 \,. \tag{2.34}$$

Unfortunately, the computation of (2.32) is very complex because of the time correlation of the fading process. To overcome this problem, we replace the continuous time model with a slotted-time model, where the UE's trajectory is observed at time epochs spaced apart by the fading coherence time T_c, given in (2.4). In this way, at each slot we can approximately assume an independent fading value. Note that the sampling time, i.e., the slot duration, varies with the UE's speed, according to (2.4). Nonetheless, the distance covered by the UE in a time slot is constant and equal to

$$\Delta_c = vT_c = \sqrt{\frac{9}{16\pi}} \frac{c}{f_c} \,. \tag{2.35}$$

In the following, we will refer to the space interval Δ_c, which represents the spatial granularity of our model, as *space slot*.

We can then define the *average trajectory capacity* $\bar{\mathcal{C}}_L$ with respect to this sampled space as

$$\bar{\mathcal{C}}_L = \frac{2}{\pi} \int_0^{\pi/2} \frac{1}{N_L} \sum_{k=1}^{N_L} \sum_{S \in \{M,F,H\}} \bar{C}_S(\mathbf{a}_k(\omega)) \mathrm{P}_S[\mathbf{a}_k(\omega)] \, \mathrm{d}\omega \,, \tag{2.36}$$

where

$$N_L = \left\lceil \frac{L}{\Delta_c} \right\rceil \tag{2.37}$$

is the total number of sample points along the trajectory, and $\mathrm{P}_S[\mathbf{a}_k(\omega)]$ is the probability that the UE is in state $S \in \{M, F, H\}$ at sample point $\mathbf{a}_k$ along its trajectory. In the next section, we describe a Markov model to compute the probabilities $\mathrm{P}_S[\mathbf{a}_k(\omega)]$.

We point out that the Markov analysis in the following section and the subsequent handover policy remain valid even with a more general propagation model than (2.3), i.e., with other random processes used to describe the fading effect. The crucial aspect is that the independence of successive fading samples must be ensured by choosing a proper sampling period T_c for that channel model. The Rayleigh fading distribution used in (2.3) allows a semi-closed form expression for the probabilities $\mathrm{P}_S[\mathbf{a}_k(\omega)]$, whereas they can be obtained through numerical methods for any other fading distribution.

2.5 Markov analysis to compute of the mean trajectory performance

In this section we model the HO process by means of a *non homogeneous* discrete time Markov Chain (MC). To begin with, we denote by N_T and N_H the number of space slots covered by the UE in time T and T_H, respectively, i.e.,

$$N_T = \left\lceil \frac{vT}{\Delta_c} \right\rceil , \quad N_H = \left\lceil \frac{vT_H}{\Delta_c} \right\rceil . \tag{2.38}$$

At every step, the UE moves along its trajectory, and the SINR changes accordingly. As explained in the previous section, the HO process is started whenever the SINR drops below a certain threshold γ_{th}. We then define M_j and F_j, with $j \in \{0, \ldots, N_T\}$, as the MC state that is entered when the UE is connected to the M-BS or F-BS, respectively, and the SINR has remained below γ_{th} for j consecutive steps. Furthermore, we define H_j and $\tilde{H}_j$, $j \in \{1, \ldots, N_H\}$, as the MC states entered when the UE performs the macro-to-femto and femto-to-macro handover, respectively.

Assume that, at step k, the MC is in state M_j. In the following step, the MC evolves from M_j to M_{j+1} if $\gamma_M(\mathbf{a}_k, kT_c) < \gamma_{th}^M$, otherwise the MC returns to M_0 since the TTT counter is reset. Conversely, if the SINR remains below threshold when the MC is in state M_{N_T}, the UE starts the HO process to the F-BS and the MC enters state H_1. In the following N_H steps the MC deterministically crosses all the handover states H_j and ends up in state F_0, regardless

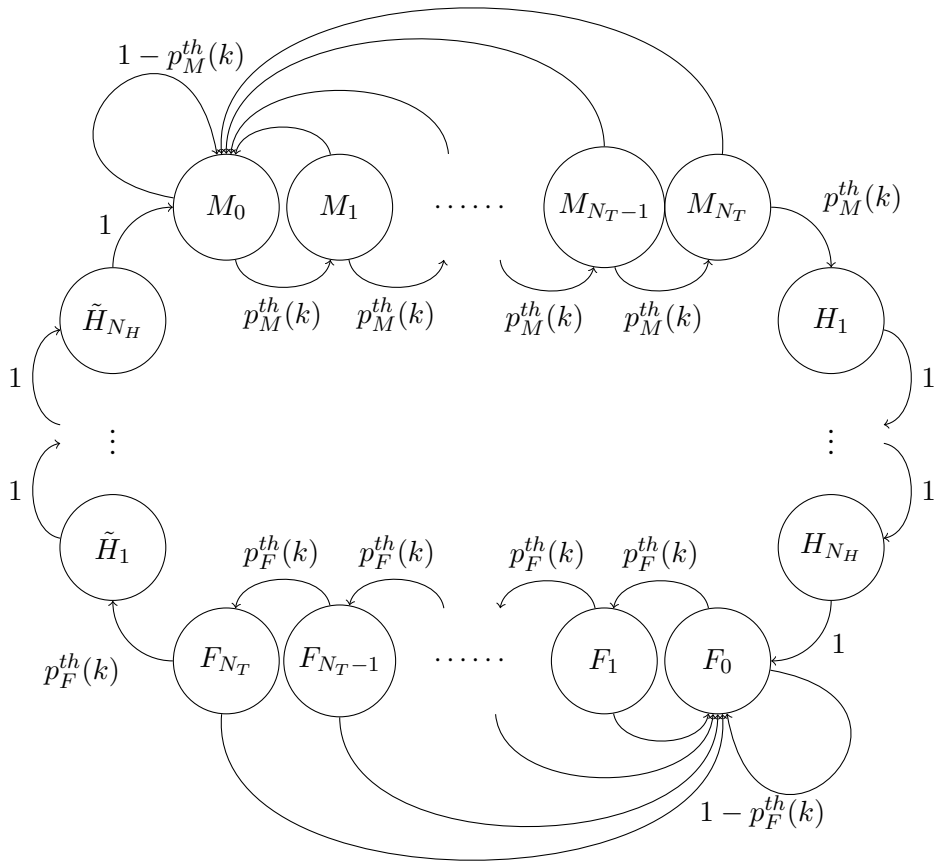

Figure 2.7. *Non homogeneous discrete time Markov chain referred to a scenario with arbitrary N_T and N_H. The transition probabilities are given by (2.40) and (2.41).*

of the channel conditions. At this point, the UE is connected to F-BS, and the evolution of the MC is conceptually identical to that seen for the M_j states.

A graphical representation of the non homogeneous discrete time MC is shown in Fig. 2.7, with the transition probabilities that will be explained below.

2.5.1 Transition probabilities and transition matrix

The cumulative distribution function of the random variable ξ_h, given in (2.7) as the ratio of two independent and identically distributed exponential random variables, is equal to

$$\mathrm{P}[\xi_h \leq x] = \frac{x}{x+1}\,, \quad x \in [0, +\infty]\,. \tag{2.39}$$

Using (2.6) and (2.39), the transition probability from state M_j to M_{j+1}, with $j \in \{0, \ldots, N_T\}$[7], at step k, is given by

$$p_M^{th}(k) = \mathrm{P}\left[\gamma_M(\mathbf{a}_k, kT_c) < \gamma_{th}^M\right] = \frac{\gamma_{th}^M}{\gamma_{th}^M + \bar{\gamma}_M(\mathbf{a}_k)}\,. \tag{2.40}$$

Similarly, the transition probability from F_j to F_{j+1} is equal to

$$p_F^{th}(k) = \mathrm{P}\left[\gamma_F(\mathbf{a}_k, kT_c) < \gamma_{th}^F\right] = \frac{\gamma_{th}^F}{\gamma_{th}^F + \bar{\gamma}_F(\mathbf{a}_k)}\,. \tag{2.41}$$

Note that (2.40) and (2.41) vary along the UE trajectory because of the pathloss, so that the MC is indeed non-homogeneous.

Without loss of generality, we can arrange the states according to the order $\{M_j\}$, $\{H_j\}$, $\{F_j\}$, and $\{\tilde{H}_j\}$, and in increasing order of the index j within the same set of states. The system transition matrix $\mathbf{P}(k)$ at the k-th step can then be expressed with the following sub block structure

$$\mathbf{P}(k) = \begin{bmatrix} \mathbf{M}(k) & \mathbf{V}_M^H(k) & \varnothing & \varnothing \\ \varnothing & \mathbf{H}(k) & \mathbf{V}_H^F(k) & \varnothing \\ \varnothing & \varnothing & \mathbf{F}(k) & \mathbf{V}_F^{\tilde{H}}(k) \\ \mathbf{V}_{\tilde{H}}^M(k) & \varnothing & \varnothing & \tilde{\mathbf{H}}(k) \end{bmatrix} \tag{2.42}$$

where the submatrices $\mathbf{M}(k)$, $\mathbf{F}(k)$, $\mathbf{H}(k)$, and $\tilde{\mathbf{H}}(k)$ are the square transition matrices within the sets $\{M_j\}$, $\{F_j\}$, $\{H_j\}$, and $\{\tilde{H}_j\}$, respectively, while $\mathbf{V}_X^Y(k)$ are the rectangular transition matrices from set X to set Y. The elements of other blocks, represented by the symbol $\varnothing$, are all equal to 0. From the previous analysis, $\mathbf{M}(k)$ is given by

$$\mathbf{M}(k) = \begin{bmatrix} 1 - p_M^{th}(k) & p_M^{th}(k) & 0 & \cdots & 0 \\ 1 - p_M^{th}(k) & 0 & p_M^{th}(k) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 - p_M^{th}(k) & 0 & 0 & \cdots & p_M^{th}(k) \\ 1 - p_M^{th}(k) & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{2.43}$$

[7] With $M_{N_T+1} \equiv H_1$.

$\mathbf{F}(k)$ is the same as $\mathbf{M}(k)$ with $p_F^{th}(k)$ in place of $p_M^{th}(k)$, while

$$\mathbf{H}(k) = \tilde{\mathbf{H}}(k) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & 1 \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix} . \tag{2.44}$$

Finally,

$$\mathbf{V}_H^F(k) = \mathbf{V}_{\tilde{H}}^M(k) = \begin{bmatrix} \varnothing & \varnothing \\ 1 & \varnothing \end{bmatrix} , \tag{2.45}$$

and

$$\mathbf{V}_M^H(k) = \begin{bmatrix} \varnothing & \varnothing \\ p_M^{th}(k) & \varnothing \end{bmatrix} , \mathbf{V}_F^{\tilde{H}}(k) = \begin{bmatrix} \varnothing & \varnothing \\ p_F^{th}(k) & \varnothing \end{bmatrix} . \tag{2.46}$$

The state probability vector $\mathbf{p}(k)$ at the k-th step is given by

$$\mathbf{p}(k) = \mathbf{p}(0) \prod_{i=0}^{k-1} \mathbf{P}(i) , \tag{2.47}$$

where $\mathbf{p}(0)$ is the state probability vector at the starting point of the UE trajectory, and $\mathbf{P}(i)$ is the transition matrix defined at the i-th step along the UE trajectory. Assuming that the UE starts its path when connected to the M-BS, we set the initial probabilities to 1 for M_0 and 0 for all the other states, so that

$$\mathbf{p}(0) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} . \tag{2.48}$$

We can then compute the probability that the UE is in state $S \in \{M, F, H\}$ at any given point $\mathbf{a}_k$, $k \in \{1, \ldots, N_L\}$, as the sum of the probabilities of the states $\{M_j\}$, $\{F_j\}$, and $\{H_j\} \cup \{\tilde{H}_j\}$, respectively, at step k, i.e.,

$$\mathrm{P}_S[\mathbf{a}_k] = \sum_{i \in \{S_j\}} p_i(k) , \tag{2.49}$$

where $p_i(k)$ is the i-th entry of the state probability vector (2.47).

2.6 Handover Decision accounting for Cell Load

In this section we consider the handover decision problem when macro and femtocells are partially loaded. In this case, handing over towards the BS with the strongest RSRP may

actually yield poorer performance because of the traffic load of the new cell. As in [50], we assume that the BSs include an indication of their current traffic load in the pilot signals, so that the UEs know the average fraction of available resources for each surrounding cell. This information shall then be considered in the HO strategy, in order to select the cell with the best tradeoff between signal quality and traffic load.

Let $\lambda_S \in [0,1]$, $S \in \{M, F\}$, denote the fraction of available resources in the cell served by S-BS. Although our model can accommodate any other scaling law, for the sake of simplicity we assume that the average performance experienced by a UE when connected to such a BS will be simply proportional to λ_S. We hence define the load-scaled average capacity of the UE in state $S \in \{M, F\}$ as follows

$$\bar{C}_S^{load}(\mathbf{a}_k) = \lambda_S \bar{C}_S(\mathbf{a}_k) = \lambda_S \log_2(\bar{\gamma}_S(\mathbf{a}_k)) \frac{\bar{\gamma}_S(\mathbf{a}_k)}{\bar{\gamma}_S(\mathbf{a}_k) - 1}\,, \tag{2.50}$$

while, as usual, we assume zero capacity during handover, i.e.,

$$\bar{C}_H^{load}(\mathbf{a}_k) = 0\,. \tag{2.51}$$

Accordingly, the average load-scaled capacity $\bar{\mathcal{C}}_L^{load}$ along the UE trajectory is given by

$$\bar{\mathcal{C}}_L^{load} = \frac{2}{\pi} \int_0^{\pi/2} \frac{1}{N_L} \sum_{k=1}^{N_L} \sum_{S \in \{M,F,H\}} \bar{C}_S^{load}(\mathbf{a}_k(\omega)) \mathrm{P}_S^{load}[\mathbf{a}_k(\omega)]\, \mathrm{d}\omega \tag{2.52}$$

where $\mathrm{P}_S^{load}[\mathbf{a}_k(\omega)]$ is the probability that at point $\mathbf{a}_k$ the UE is in state $S \in \{M, F, H\}$. Clearly, this probability depends on the HO policy, which shall be adjusted to account for the load conditions of the cells.

A simple way to reach this goal, with minimal impact on the HO mechanism, is to maintain the standard SINR-based HO procedure considered in the previous section, and acting on the Cell Individual Offset (CIO) of the cells, which shall be modified to account for the different traffic loads. This is equivalent to defining, for each cell S, a threshold $\gamma_{th}^{S,load}$ that depends on the current traffic loads of the macro and femtocells, respectively.

The choice of the thresholds determines the characteristics of the load-aware HO algorithm. A reasonable approach is to adapt the threshold to the cell loads in such a way that the relative performance gain experienced by the UE when changing BS is constant. Now, averaging over the fading phenomena and assuming both macro and femtocells are unloaded ($\lambda_M = \lambda_F = 1$), the HO from M-BS to F-BS is triggered when the SINR drops below

the threshold γ_{th}^{M}. According to (2.33), the ratio between the average capacity of the UE in states M and F at this threshold-crossing point $\mathbf{a}_{k^*}$ is given by

$$\frac{\bar{C}_M(\mathbf{a}_{k^*})}{\bar{C}_F(\mathbf{a}_{k^*})} = \frac{\log_2(\gamma_{th}^{M})\frac{\gamma_{th}^{M}}{\gamma_{th}^{M}-1}}{\log_2(1/\gamma_{th}^{M})\frac{1/\gamma_{th}^{M}}{1/\gamma_{th}^{M}-1}} = \gamma_{th}^{M}\,, \tag{2.53}$$

where $\bar{\gamma}_M(\mathbf{a}_{k^*}) = \gamma_{th}^{M}$ and $\bar{\gamma}_F(\mathbf{a}_{k^*}) = 1/\gamma_{th}^{M}$. We can then set $\gamma_{th}^{M,load}$ in such a way that the ratio between the load-scaled capacities given by (2.50) at the new threshold-crossing point $\mathbf{a}_{k^*}^{load}$ is still equal to γ_{th}^{M}, i.e.,

$$\frac{\bar{C}_M^{load}(\mathbf{a}_{k^*}^{load})}{\bar{C}_F^{load}(\mathbf{a}_{k^*}^{load})} = \gamma_{th}^{M}\,. \tag{2.54}$$

where $\bar{\gamma}_M(\mathbf{a}_{k^*}^{load}) = \gamma_{th}^{M,load}$ and $\bar{\gamma}_F(\mathbf{a}_{k^*}^{load}) = 1/\gamma_{th}^{M,load}$. Using (2.50) into (2.54) we finally get

$$\gamma_{th}^{M,load} = \gamma_{th}^{M}\frac{\lambda_F}{\lambda_M}\,. \tag{2.55}$$

Repeating the same reasoning for the femto-to-macro handover, we get

$$\gamma_{th}^{F,load} = \gamma_{th}^{F}\frac{\lambda_M}{\lambda_F}\,. \tag{2.56}$$

Using $\gamma_{th}^{S,load}$ in place of γ_{th}^{S} in (2.40) and (2.41), we can then resort to the MC model described in the previous section to compute the average trajectory performance achieved by the load-aware HO policy. The model can then be utilized to investigate the optimal choice of the TTT parameter, as will be explained in the next section.

2.7 Context-Aware HO Policy (CAHP)

The mathematical model developed in Sec. 2.4–2.6 can be used to derive a *Context-Aware HO Policy* (CAHP). The context parameters that the model is built upon consist of the transmit powers of the BSs (Γ_M^{tx} and Γ_F^{tx}), the path loss coefficients (which determine the distance-dependent path gains $g_M(\mathbf{a})$ and $g_F(\mathbf{a})$), the inter-BS distance d_{MF}, the carrier frequency f_c, and the UE speed v. In addition, the traffic load of the cells can be considered for the traffic-aware CAHP. Given these parameters, it is then possible to use the models (2.36) and (2.52) to find the value TTT that maximizes the estimated average performance experienced by the UE when crossing the area. The CAHP, hence, consists in using the optimal TTT value for the current context parameters, which are supposed to be either known by the UE or estimated from the RSRP received from the different BSs. In fact, pilot signals can carry all the

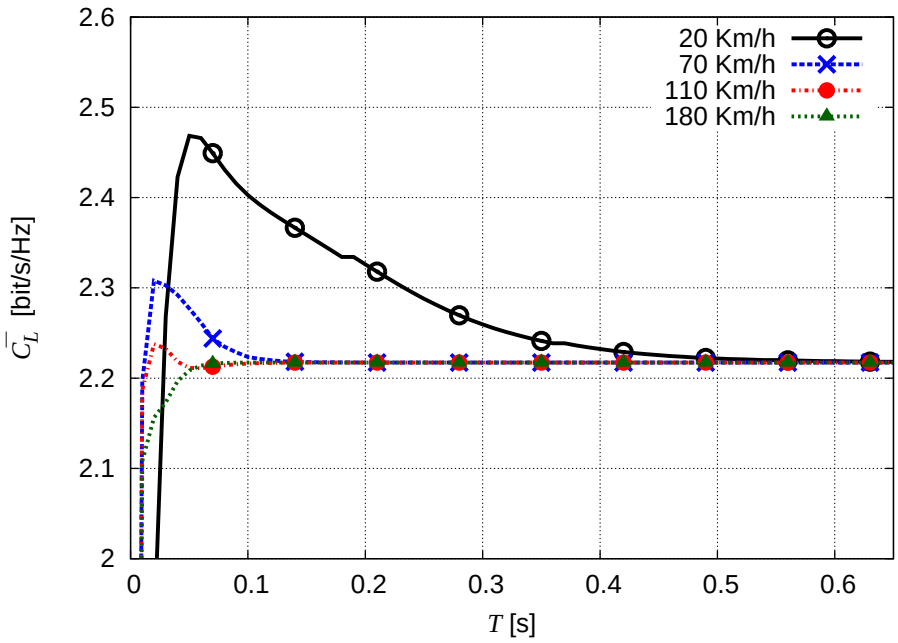

Figure 2.8. *Analytical average trajectory capacity obtained for different speeds, as a function of the TTT.*

necessary information, such as the pathloss exponent used in the propagation model and the cell load conditions, while the UE speed can be accurately obtained from the UE itself, with standard GPS-based systems provided by current devices.

In the remainder of this section we investigate the average UE capacity (2.36) when varying the context parameters, in order to gain insight on the shape of the CAHP when the cell traffic load is neglected. In the following section, we compare by simulation the performance of our CAHP against the standard handover process using static TTT values (FIX) and we extend the analysis to the model described in Sec. 2.6, where the load of the two cells is considered.

We assume a scenario composed by a M-BS with transmission power of 46 dBm and a F-BS with transmission power of 24 dBm [48]. The BSs are placed 500 m apart. Furthermore, we set $T_H = 200$ ms, $\gamma_{th}^M = \gamma_{th}^F = 1$ dB, while T is varied with a granularity of 10 ms.

Fig. 2.8 shows the analytical average capacity $\bar{C}_L$ given by (2.36) for different speeds, as a function of T. We note that the curves show a similar trend for all speed values. The sharp capacity drop for low T values is due to the ping-pong effect, which is indeed alleviated when using longer T values. In particular, the longer the channel coherence time (i.e., the lower the speed v), the larger the T required to avoid the ping pong effect, as expected.

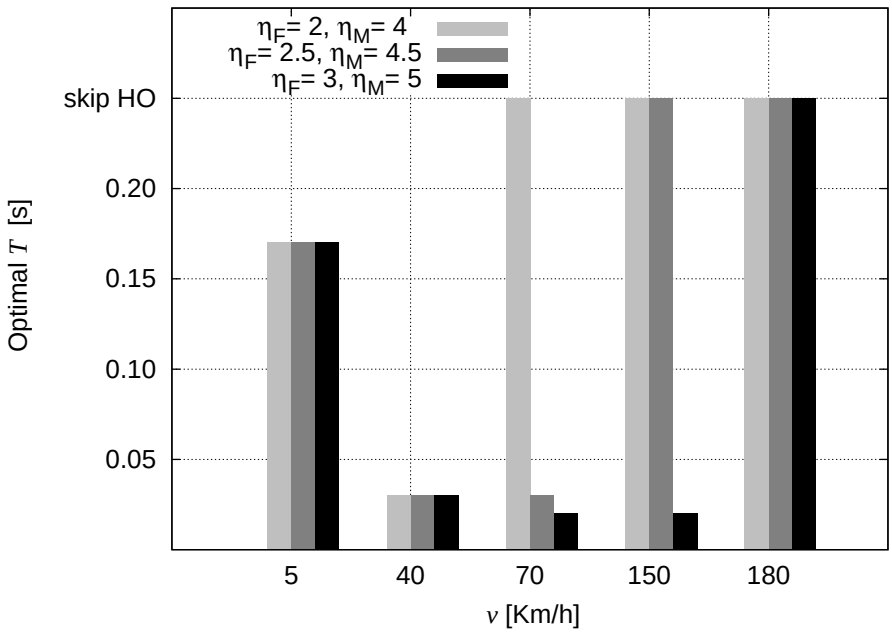

Figure 2.9. *Optimal T for different UE speeds v and channel parameters according to the CAHP approach.*

For high T values, all curves reach an asymptotic value that corresponds to the average trajectory capacity achievable when handover is not performed. The optimal T shall then trade off between the risk of ping-pong effect and the HO delay. Note that, for very high UE speeds, the maximum capacity corresponds to the asymptotic capacity. In this case, the optimal policy simply consists in always avoiding the HO, since the performance loss incurred during the HO process is never compensated by the capacity gain obtained by connecting to the F-BS.

Fig. 2.9 shows the optimal T values obtained from the analytical model for different speeds and scenarios. In practice, we vary the pathloss coefficients of the macro and femto BSs to change the channel profile and the femtocell coverage area, which is "small" for $\eta_F = 2, \eta_M = 4$ (radius of 9 m, left most bar), "medium", for $\eta_F = 2.5, \eta_M = 4.5$ (radius of 11 m, middle bar), and "large", for $\eta_F = 3, \eta_M = 5$ (radius of 13 m, right most bar). As predictable, the speed threshold above which the optimal policy is to skip HO depends on the femtocell range. In particular, for large cells, the losses due the HO are balanced by the higher capacity obtained by connecting to the F-BS. Therefore, skipping HO is convenient only when the UE speed is quite high. For lower speeds, instead, the optimal T is the minimum value to avoid ping-pong events due to fast fading and, hence, only depends on

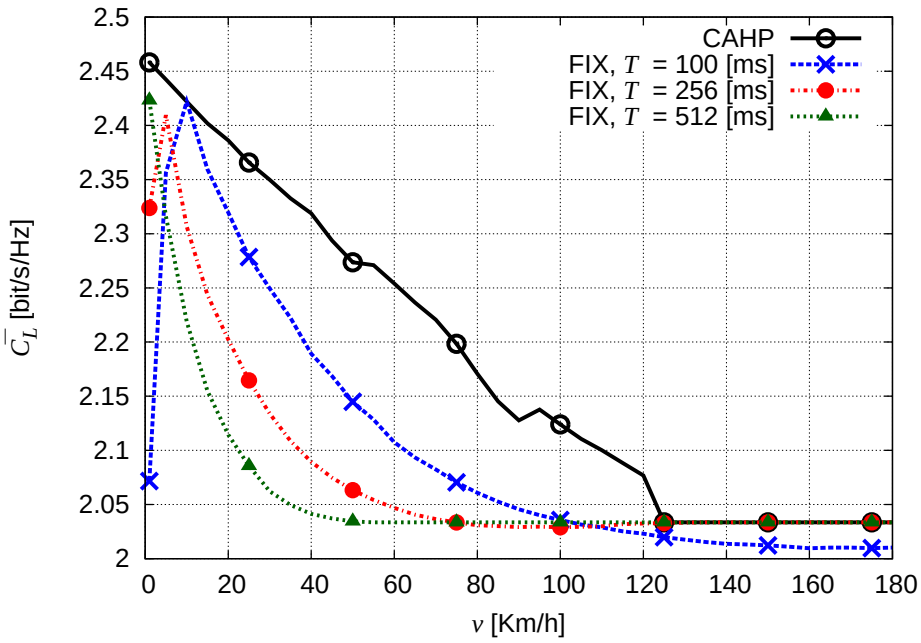

Figure 2.10. *Average capacity trajectory obtained with different approaches, as a function of the UE speed.*

the channel coherence time that, in turn, depends on the UE's speed, but is independent of the size of the cells.

2.8 Performance Evaluation

In this section we evaluate the performance achieved by the CAHP approach through Montecarlo simulations. In particular, we compare the mean capacity obtained by CAHP against the capacity of FIX policies that use constant TTT values, with $T \in \{100\,\text{ms}, 256\,\text{ms}, 512\,\text{ms}\}$, irrespective of the UE speed and of the other channel parameters. In the simulation we consider path loss coefficients $\eta_F = 2.5$ and $\eta_M = 4.5$ for F-BS and M-BS, respectively, the fast fading model presented in Sec. 2.2, and a noise level equal to $\sigma^2 = -130$ dBm, obtained assuming a total downlink bandwidth of 20 MHz and a noise power spectral density of $N = k_B \mathrm{T}_0 = -143.82$ dBW/MHz, where the noise temperature T_0 is equal to 300 K and k_B is the Boltzmann constant.[8]

Fig. 2.10 shows the average trajectory capacity obtained in the simulations. At low speeds, the performance of the FIX policy suffers from the ping-pong effect due to low T

[8]We verified that these results are essentially the same that would be obtained in the absence of noise.

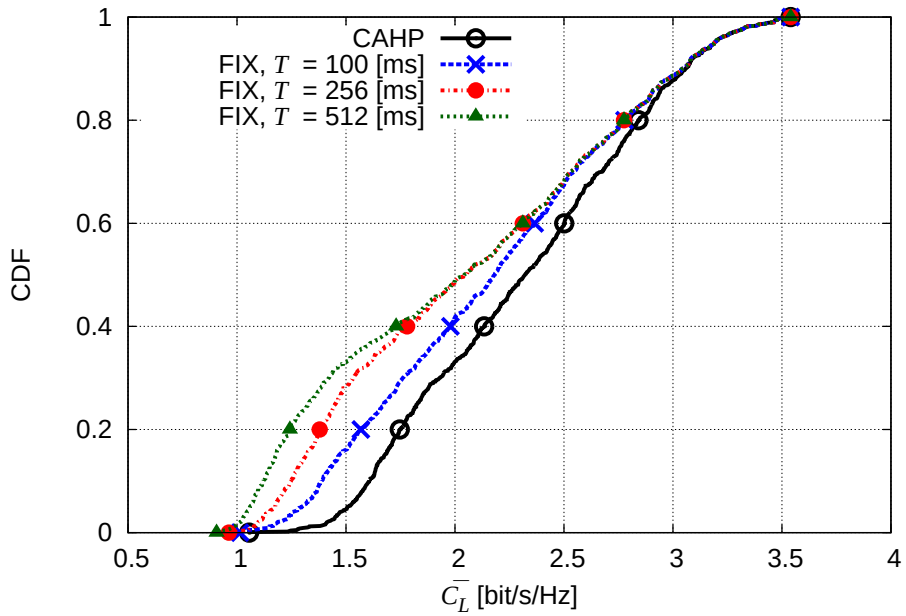

Figure 2.11. *Average trajectory capacity CDF for different approaches.*

values, while CAHP adopts a larger T that avoids HO triggering due to fast-fading fluctuations. Conversely, for higher speeds, CAHP outperforms the FIX policy by adopting sufficiently low T values to avoid the ping-pong effects, while not excessively delaying the switching to the F-BS. In particular, the higher the fixed T value, the lower the speed beyond which HO is never performed, and the higher the capacity loss compared to CAHP that, instead, performs a handover. We note that, at high speeds, all curves asymptotically converge to the same value corresponding, as in the analytical model, to the average trajectory capacity achieved when the UE remains always connected to the M-BS. The optimal HO policy consists therefore in not performing the handover to the F-BS, to avoid the loss due to two zero-capacity T_H intervals in a short time. In this case, all policies with sufficiently large T obtain the same results. Note that the asymptotic capacity given by simulations slightly differs from that given by the Markov model, as reported in Fig. 2.8. This small discrepancy is likely due to the simplifying assumption of the analytical model, which considers a perfectly homogeneous scenario around the femtocell center $\mathbf{c}$. The simulations, instead, consider the actual location of both BSs and the actual power received at any given point by each of them.

Fig. 2.11 describes the cumulative distribution function (CDF) of the average trajectory

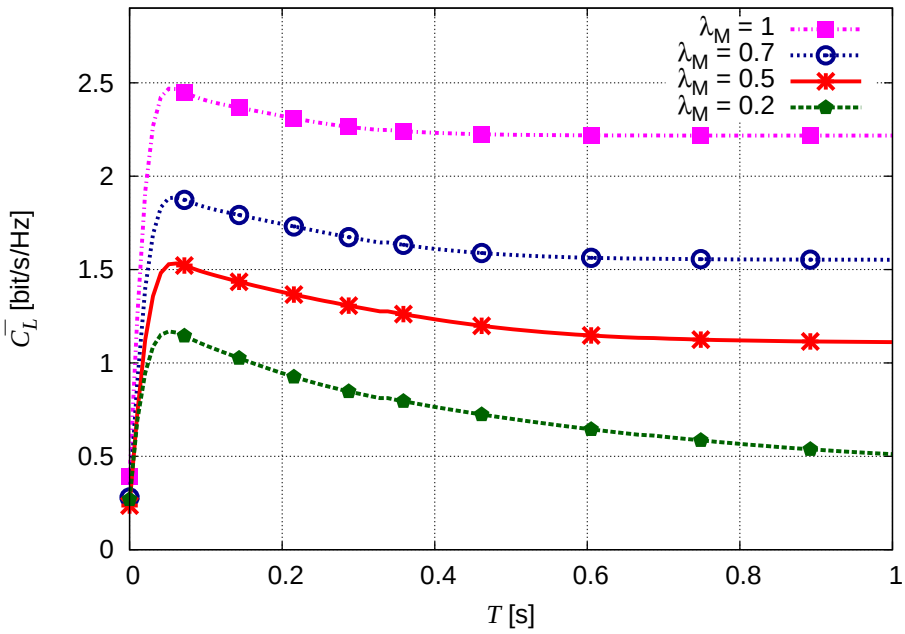

Figure 2.12. *Analytical average trajectory capacity obtained for different load conditions, as a function of* T*, with* $v = 20$ Km/h.

capacity for a UE speed of $v = 40$ Km/h. We note that the improvement provided by CAHP is concentrated in the lower part of the CDF. These values correspond to the trajectories that cross the femtocell area close to its center, i.e., to the location of the F-BS. In this region, a small T makes it possible to exploit the signal from the F-BS and to gain up to 50% in capacity in comparison with the case with larger T. On the contrary, the higher part of the CDF corresponds to trajectories that cross the femtocell far from the center, so that the average trajectory capacity is basically unaffected by T because HO is skipped in most cases.

The above results have been obtained by assuming that both the macro and the femtocell were unloaded. In the following we instead consider the case where the capacity of the cells is partially taken by other users. The pathloss coefficients from M-BS and F-BS are fixed to 4.5 and 2.5, respectively. Fig. 2.12 shows the analytical average trajectory capacity (2.52) as function of T, and with UE's speed $v = 20$ Km/h, when varying the load factor λ_M of the macrocell in the set $\lambda_M \in \{0.2, 0.5, 0.7, 1\}$, while keeping the femtocell unloaded ($\lambda_F = 1$). We can observe that the curves in Fig. 2.12 have the same shape, but are scaled according to λ_M. In particular, the asymptotic capacity scales proportionally to λ_M. In fact, when T is large enough, the UE does not perform any handover and remains always connected

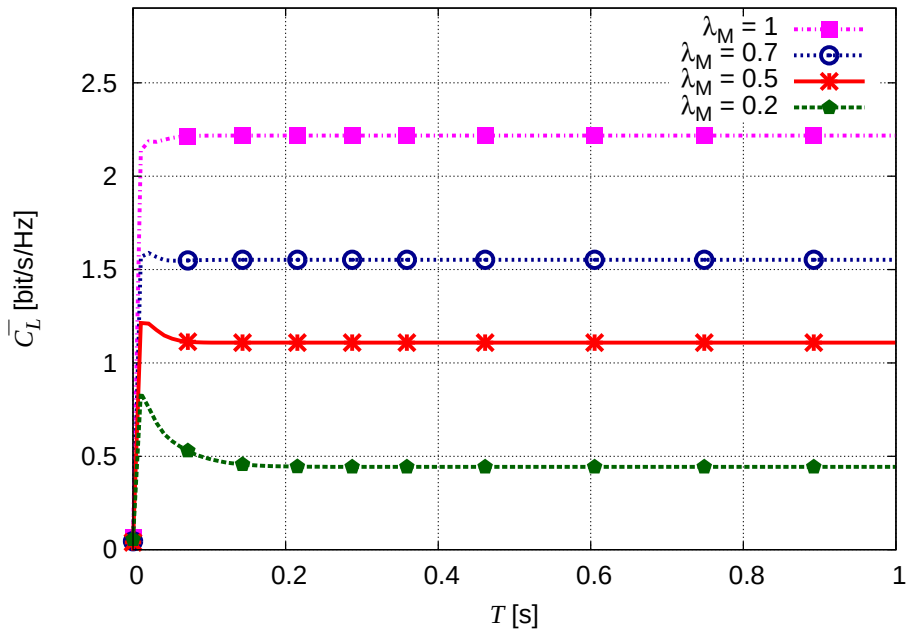

Figure 2.13. *Analytical average trajectory capacity obtained for different load conditions, as a function of* T*, with* $v = 150$ Km/h.

to the macrocell, and its resulting average trajectory capacity equals that of the macrocell, which is scaled by a factor λ_M with respect to the unloaded case. We also observe that the T value that maximizes the average trajectory capacity is the same for every load condition. The situation however changes for higher UE speed, as can be seen from Fig. 2.13 which reports the average capacity of the UE when varying T, with $v = 150$ Km/h. Here, CAHP encourages the UE to switch to the femtocell for highly loaded macrocells ($\lambda_M = 0.2, 0.5$), while it avoids the handover when the macrocell is unloaded. This confirms the intuition that the threshold speed increases with the load of the macrocell.

Figs. 2.14 and 2.15 show the average trajectory capacity obtained through simulations when fixing $\lambda_F = 1$ and setting λ_M equal to 0.2 and 0.7, respectively. In order to quantify the performance achieved by CAHP, we show also the capacity upper bound (Opt) computed in Sec. 2.9, that represents the best achievable performance for every user trajectory. Note that the computation of the optimal strategy requires to know in advance the fast fading gains at each point along the UE's trajectory and, hence, it is infeasible in practical scenarios. As in the previous case, we compare the performance achieved by the CAHP policy with two TTT-fixed policies, where the cell loads are not considered and T is set to 100 ms and 50 ms,

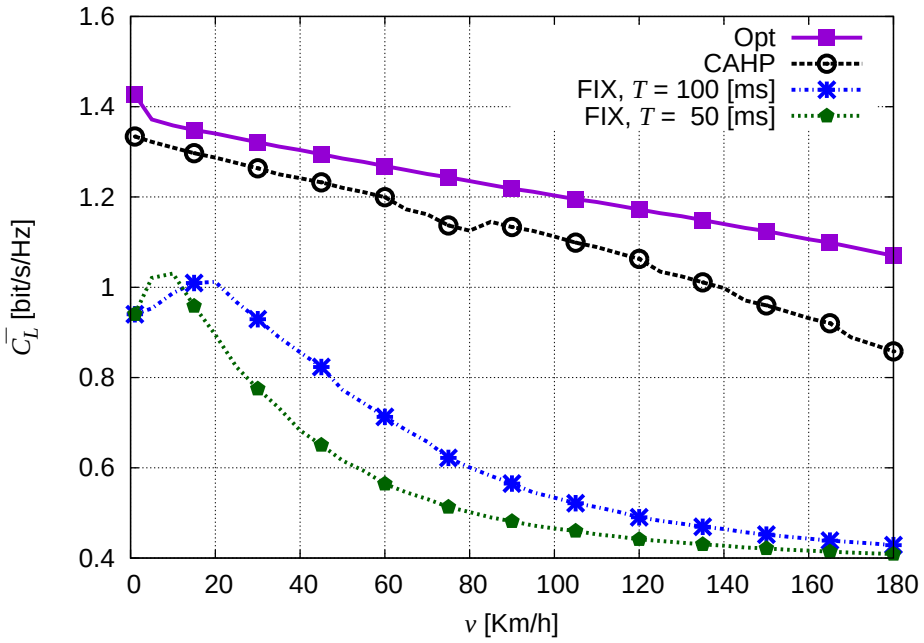

Figure 2.14. *Average trajectory capacity obtained with different approaches with* $\lambda_M = 0.2$.

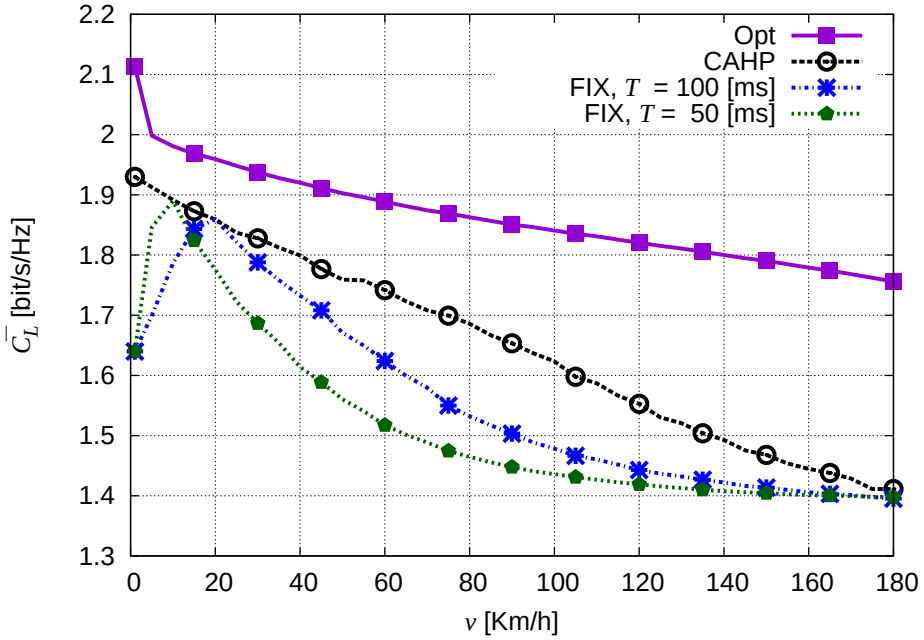

Figure 2.15. *Average trajectory capacity obtained with different approaches with* $\lambda_M = 0.7$.

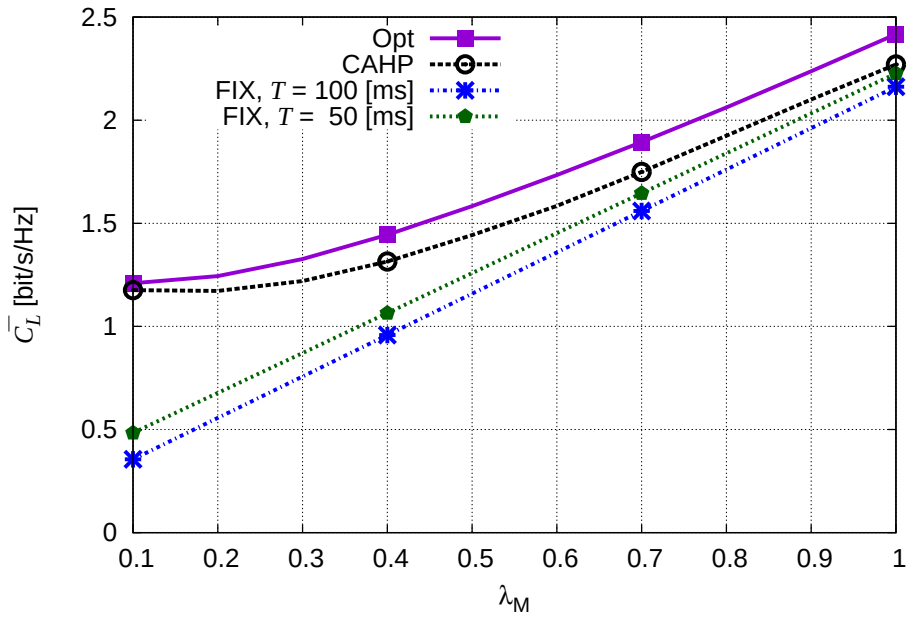

Figure 2.16. *Average trajectory capacity obtained with different approaches for* $v = 60\ \mathrm{Km/h}$ *and varying* λ_M *from* 0.1 *to* 1.

respectively. As in Fig. 2.10, the CAHP approach achieves a substantial gain in comparison with the TTT-fixed policies for all the considered speeds. We notice that, since the capacity penalty due to T_H is larger at high speeds, the gap with the Opt policy increases with the users velocity. Moreover, the gain provided by the CAHP policy grows when the cell load is unbalanced.

This trend is further analyzed in Fig. 2.16. In this simulation we set $v = 60\ \mathrm{Km/h}$, while λ_M is varied from 0.1 to 1 and $\lambda_F = 1$. As expected, the average trajectory capacity increases when the macrocell is unloaded since HO is performed less frequently because the macrocell provides good enough performance. When the load at the macrocell increases, the gap between the CAHP and the TTT-fixed policies increases. The CAHP gain is due to the capability of the CAHP approach to tune the TTT considering the cell loads. In particular, when the load at the macrocell is very high, the CAHP policy achieves more than 100% performance improvement with respect to the TTT-fixed policies.

2.9 Upper Bound Analysis of the Handover process

Handover in HetNets is an interesting research challenge that has attracted considerable attention in the last years, producing a variety of handover schemes that differ in the considered assumptions and target user/network utility functions. Therefore, such schemes are hardly comparable, and the remaining space for further optimization is quite unclear. In this section, we propose a general framework to derive the limiting performance of handover in HetNets. Our scheme assumes non-causal knowledge of the channel samples and computes the optimal handover strategy as a function of the user's speed, cell size and load conditions. The proposed framework is useful not only to determine the margin of improvement of the existing handover schemes, but also to provide a comparative performance analysis among them.

The addressed scenario consists of a set $\{B_0, B_1, \ldots, B_N\}$ of BSs, whose locations are assumed to be known. We then target a mobile UE that crosses the area along a certain trajectory. Fig. 2.18 gives an example of the considered scenario. We denote by $\Gamma_i(\mathbf{a}, t)$ the RSRP that the UE collected from B_i at time t, when it was in position $\mathbf{a}$ along the trajectory. Assuming a pathloss plus fading propagation model, in the downlink channel we then have

$$\Gamma_i(\mathbf{a}, t) = \Gamma_i^{tx} g_i(\mathbf{a}) \alpha_i(t), \quad i \in \{0, \ldots, N\}\,; \tag{2.57}$$

where Γ_i^{tx} is the transmit power of B_i, $g_i(\mathbf{a})$ is the pathloss from B_i to point $\mathbf{a}$, and $\alpha_i(t)$ is the fast-fading channel gain at time t.[9]

Assuming exact knowledge of the UE trajectory and of the BS positions, we can hence determine the average performance experienced by the UE when crossing the area, for a given HO strategy. For analytical tractability, however, it is convenient to consider a discrete version of the problem that is obtained by sampling the process $\Gamma_i(\mathbf{a}, t)$ with a time step T_c that makes it possible to neglect the fading correlation. For instance, for a given UE speed v, and assuming Rayleigh fading, T_c may be set equal to the channel coherence time, given in (2.4).

The sampled version of the RSRP process, hence, can be written as

$$\Gamma_i(\mathbf{a}_k, \tau_k) = \Gamma_i^{tx} g_i(\mathbf{a}_k) \alpha_i(\tau_k)\,, \tag{2.58}$$

[9] Fading is assumed unknown. However, the proposed framework can also be used to determine the achievable HO performance under exact and non-causal knowledge of the channel state information along the trajectory, as done at the end of the letter.

where $\mathbf{a}_k$ and τ_k are the kth sampling points along the UE trajectory and in time, respectively. Assuming the UE is served by B_i at this sampling point, the Signal-to-Interference-Ratio (SIR)[10] experienced by the UE can be expressed as

$$\gamma_i(\mathbf{a}_k, \tau_k) = \frac{\Gamma_i(\mathbf{a}_k, \tau_k)}{\sum_{j \neq i} \Gamma_j(\mathbf{a}_k, \tau_k)} \,. \tag{2.59}$$

At each point along its trajectory, the UE can either be served by a certain BS, or performing HO towards another BS. We assume that the HO process takes a time T_H to be concluded, corresponding to a certain number h of sampling intervals, and that during this time the UE is not served by any BS. We hence denote by $\mathcal{B}_i$ the state of the UE when it is connected to B_i, while $\mathcal{H}_j^\ell$ indicates that the UE is at the jth step of the HO procedure to connect to B_ℓ. The set of all possible states is then denoted as[11] $\Omega = \mathcal{B} \cup \mathcal{H}$, where $\mathcal{B} = \{\mathcal{B}_i\}_{i=0,\dots,N}$ and $\mathcal{H} = \{\mathcal{H}_j^\ell\}_{i=0,\dots,N}^{\ell=1,\dots,h}$.

Denoting by K the total number of sample points along the UE trajectory, any HO strategy can be represented by a vector of K elements, $S = [s(1), \dots, s(K)]$, where $s(i) \in \Omega$ represents the state of the UE at the ith sample point. The objective, hence, is to find the policy S^* that maximizes a certain utility function over all the K points along the trajectory followed by the UE.

If the utility function can be expressed as the sum of the utility experienced by the UE at each point along the trajectory, then the optimization problem can be solved using a simple adaptation of the Viterbi algorithm. Let $\boldsymbol{\pi}_s \subset \Omega$ be the set of states from which the UE can reach state $s \in \Omega$ in one step. It is easy to realize that

$$\boldsymbol{\pi}_s = \begin{cases} \{s\} \cup H_s^h & , \quad s \in \mathcal{B} \,; \\ \mathcal{B} \setminus \{\mathcal{B}_j\} & , \quad s = H_j^1 \,; \\ H_j^{\ell-1} & , \quad s = H_j^\ell \,, \ell \neq 1 \,; \end{cases} \tag{2.60}$$

where the second row expresses the obvious fact that the UE will never start a HO process towards the same BS it is already connected to, while the third row indicates that, once started, the HO process continues for exactly h steps. We can then build the trellis diagram of depth K, where every step $k = 1, \dots, K$ corresponds to a sample along the UE trajectory, and where state q at step $k+1$ can only be reached from a state $p \in \boldsymbol{\pi}_q$ at step k. Fig. 2.17

[10] Since the HetNet scenario is interference-limited, the noise term is neglected for simplicity and without loss of generality.

[11] The symbols $\cup$ and $\setminus$ denote the set union and the set theoretic difference operations.

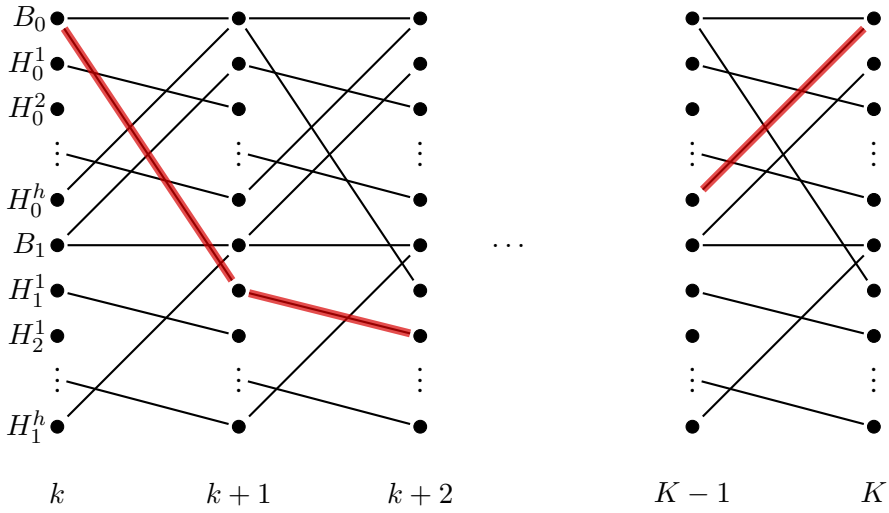

Figure 2.17. *Trellis diagram from a generic step k till the end of the UE trajectory. We assume $N = 1$, i.e., the UE can switch between two BSs.*

reports a chunk of the trellis diagram when $N = 1$, from step k to step K. Note that, while the trellis of the standard Viterbi algorithm is fully connected at every step, the precedence rules expressed in (2.60) make it possible to reduce the complexity of the algorithm from $\mathcal{O}(K(N+1)^2h^2)$ to $\mathcal{O}(K(N+1)(N+1+h))$. We remark also that the complexity can be further reduced by considering only the handover processes among neighboring cells.

Now, each link of the trellis that ends into state s at step k is assigned a certain *gain* $C_s(k)$, which only depends on the arrival state s. Following the rules of the Viterbi algorithm, and assuming the initial state of the UE is $s_0 \in \Omega$, the utility function at every step k can be expressed recursively as follows:

$$\mathcal{U}_s(k) = \max_{q \in \boldsymbol{\pi}_s} \mathcal{U}_q(k-1) + C_s(k) \;,\; \forall s \in \Omega,\; k = 1, \ldots, K \;, \tag{2.61}$$

with $U_{s_0}(0) = C_{s_0}(0)$ and $U_s(0) = 0$ for any $s \neq s_0$. Once the utility function is computed for all the possible states along the trellis, the optimal policy is obtained by starting from

$$s^*(K) = \arg \max_{s \in \Omega} \mathcal{U}_s(K) \,, \tag{2.62}$$

and going backward along the path that maximizes the utility at each step, i.e.,

$$s^*(k) = \arg \max_{q \in \boldsymbol{\pi}_{s^*(k+1)}} \mathcal{U}_q(k) \;,\; k = K-1, \ldots, 1 \,. \tag{2.63}$$

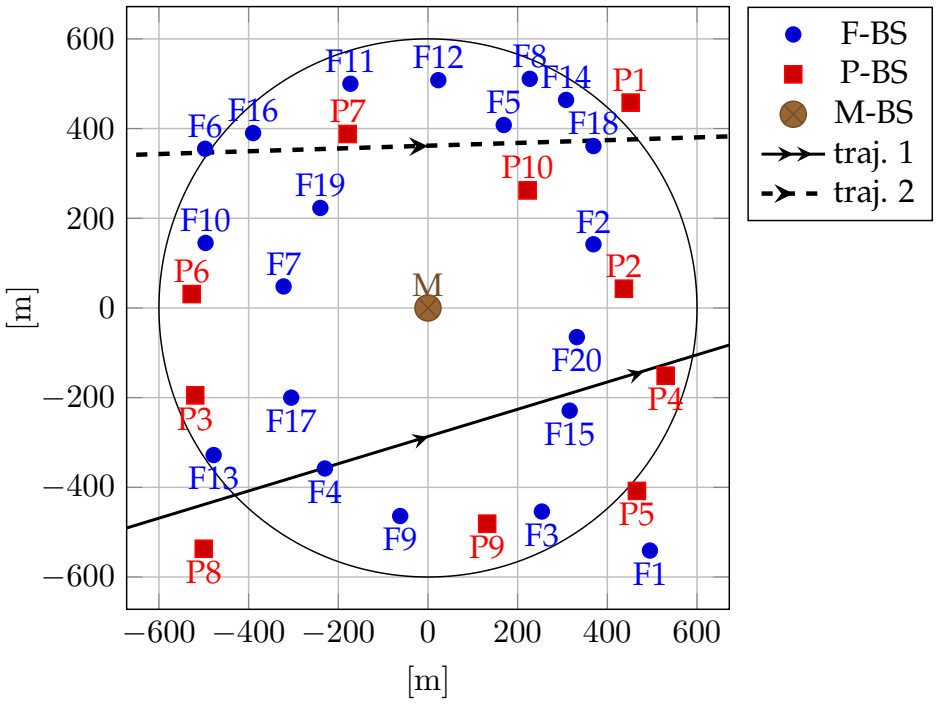

Figure 2.18. *Reference heterogeneous scenario.*

2.9.1 Performance Evaluation of the Upper Bound Analysis

In this section, we compare the performance achieved in a given scenario by some practical HO algorithms proposed in the literature, and we assess their gap with respect to the optimal HO performance obtained with the proposed model in the same scenario.

We consider the case of a single macro cell, containing N small-cell BSs. We assume the target UE follows a straight trajectory at constant speed v, and that the RSRPs are affected by Rayleigh fading, so that the coefficients $\alpha_i(t)$ are exponential random variables with unit mean. The utility function is the average Shannon capacity experienced by the UE along its trajectory. Hence, the gain in state s at time k is given by

$$C_s(k) = \begin{cases} \lambda_s \mathbb{E}\left[\log_2\left(1+\gamma_s(\mathbf{a}_k,\tau_k)\right)\right] & , \quad s \in \mathcal{B}\,; \\ 0 & , \quad s \in \mathcal{H}\,; \end{cases} \tag{2.64}$$

where $\gamma_s(\mathbf{a}_k,\tau_k)$ is defined in (2.59), while $\lambda_s \in [0,1]$ accounts for the available fraction of the cell capacity. Note that we assume zero capacity during handover, in order to reflect the performance loss incurred by the UE when switching BS. For the specific case of Rayleigh

fading, the gain (2.64) for any $s \in \mathcal{B}$ admits the closed form expression (see Appendix A.3)

$$C_s(k) = \lambda_s \sum_{i \in \mathcal{B} \setminus s} \frac{\psi_{s,i}(\mathbf{a}_k)}{1 - \frac{\bar{\Gamma}_i(\mathbf{a}_k)}{\bar{\Gamma}_s(\mathbf{a}_k)}} \log_2 \frac{\bar{\Gamma}_s(\mathbf{a}_k)}{\bar{\Gamma}_i(\mathbf{a}_k)}, \tag{2.65}$$

where

$$\psi_{s,i}(\mathbf{a}_k) = \frac{1}{\prod_{j \in \mathcal{B} \setminus \{s,i\}} \left(1 - \frac{\bar{\Gamma}_j(\mathbf{a}_k)}{\bar{\Gamma}_i(\mathbf{a}_k)}\right)} \tag{2.66}$$

and $\bar{\Gamma}_i(\mathbf{a}_k) = \Gamma_i^{tx} g_i(\mathbf{a}_k)$ is the received power averaged over the fading process. Using (2.65) into (2.61) we can finally determine the optimal HO policy through the algorithm described in the previous section.

Fig. 2.18 shows a test scenario, where 10 pico BSs and 20 femto BSs are randomly deployed within a macro cell coverage area of radius $R = 600$ m. The trajectory followed by the UE is shown in solid line (traj. 1). The powers transmitted by the three-tier cells, Macro, Pico, and Femto, are $\{P_M^{tx}, P_P^{tx}, P_F^{tx}\} = \{46, 30, 24\}$ dBm, as in [48], while the pathloss coefficients are $\{\eta_M, \eta_P, \eta_F\} = \{4.5, 2.5, 2.5\}$. In this scenario, small cells are unloaded, i.e., $\lambda_P = \lambda_F = 1$, while for the macro cell we consider two cases, with $\lambda_M = 0.2$ and $\lambda_M = 1$, respectively.

Fig. 2.19 shows the average RSRP for the macro BS M, and for the BSs that are close to the trajectory of the UE, namely the pico cells $\{\mathrm{P8}, \mathrm{P4}\}$ and femto cells $\{\mathrm{F4}, \mathrm{F15}\}$. In addition, the figure shows the average RSRP experienced by the UE when performing the optimal HO strategy in the case the macro cell is unloaded (Opt1), and heavily loaded (Opt2). As can be seen, in the second case the optimal HO strategy (thick red solid line) favors the connection to the closest (unloaded) BSs (including $\mathrm{P8}$), prolonging the permanence time in the femto and pico cell with respect to the optimal strategy when the macro cell is unloaded.

2.9.2 Simulation results of the Upper Bound Analysis

To gain insight on the room available for improvement in the design of HO procedures, we have simulated some HO algorithms found in the literature in a realistic scenario with 9 macro cells placed on a grid network and 75 pico cells and 145 femto cells randomly deployed at the macro cell edges. The fraction of available cell capacity for the macro, pico and femto cells are $\{\lambda_M, \lambda_P, \lambda_F\} = \{0.5, 0.8, 1\}$, respectively, while the transmitted powers and

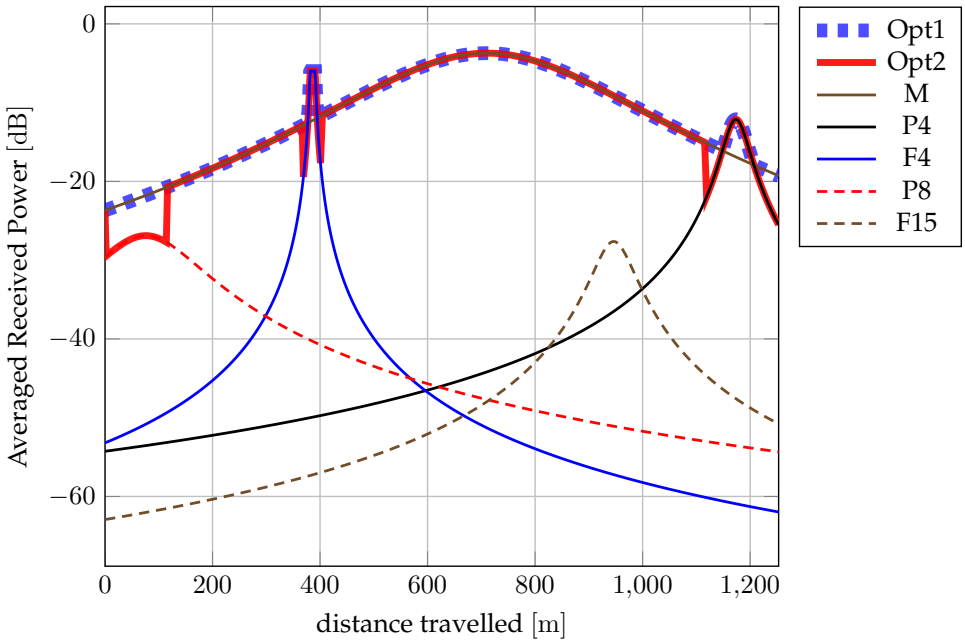

Figure 2.19. *Power profiles from the neighboring BSs along the UE trajectory 1, with speed* $v = 40\,\mathrm{Km/h}$. *The optimal policies are shown when* $\lambda_M = 1$ *(Opt1) and* $\lambda_M = 0.2$ *(Opt2).*

pathloss coefficients are the same as before. We generate random trajectories that cross the network area of size $3 \times 3.6\ \mathrm{Km}^2$.

The first HO algorithm considered in the comparison is the *Travel Distance Prediction* (*TravelDistPred*) [51], where the predicted distance within the cell coverage area is computed as soon as the RSRP of the target cell is higher than that of the serving cell. If the expected distance is higher than $2/3$ of the target cell radius, the HO is performed, otherwise it is avoided. The Time-To-Trigger (TTT) parameter, after which the HO is started, is set to $T = 10\Delta_c/v$, where v is the UE speed, and Δ_c is a fixed parameter. The second algorithm is the *Speed and Tier dependent policy* (*SpeedTier*) [37], where different TTTs are chosen according to the UE speed level (normal, medium, high) and the pair serving-target cell tiers (macro-to-macro, macro-to-small, small-to-macro, small-to-small). The third considered algorithm is the *Context-Aware Handover Policy* (*CAHP*) [15], described in Sec. 2.4–2.7, where the TTT is optimized according to UE speed and cell power profiles, while the traffic load is taken into account by properly adapting the hysteresis margin. Finally, we consider the optimal theoretical policy described in this paper (*Opt*).

In Fig. 2.20 we plot the relative capacity gap G_a of the considered algorithm a with re-

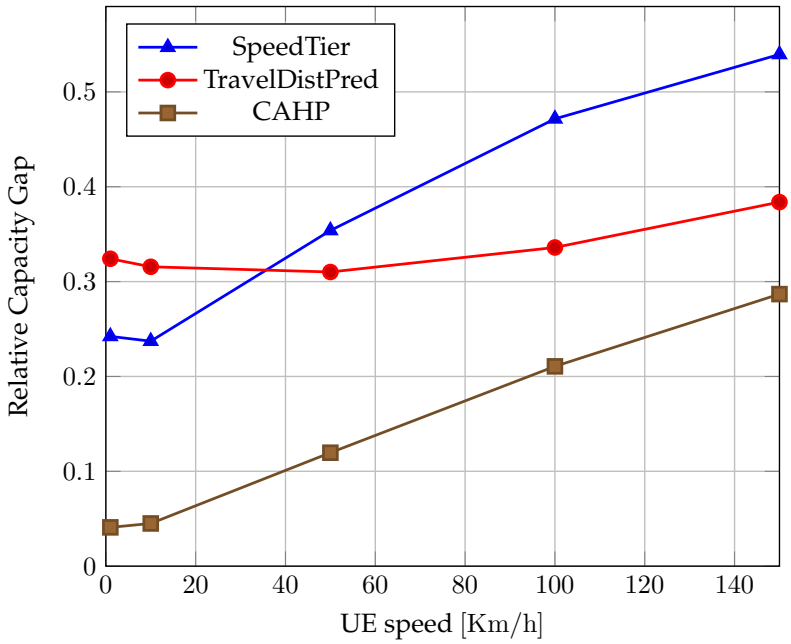

Figure 2.20. *Trajectory average capacity according to different HO policies.*

spect to the optimal policy as

$$G_a = \frac{\sum_{k\in\mathcal{K}_a}[C_{opt}(k) - C_a(k)]}{\sum_{k\in\mathcal{K}_a} C_{opt}(k)}, \tag{2.67}$$

where $C_{opt}(k)$ and $C_a(k)$ are the capacities at point k of the optimal policy and of one of the handover algorithms described above, while $\mathcal{K}_a$ is the set of points along the UE trajectory where $C_{opt}(k) \neq C_a(k)$. We can observe how the performance of *SpeedTier* and *TravelDistPred* intersect when varying v, while *CAHP*, that takes into account different context parameters (including cell loads), achieves higher performance, closer to the optimal, for different UE speeds.

Finally, we use our mathematical model to gain insight on the performance that could be achieved by knowing the exact value of the RSRP (including the fading terms) at each point of the trajectory. To this end, we ran 1000 independent simulations of a UE crossing the macro cell along trajectory 2 in Fig. 2.18 and, for each realization, we computed the optimal HO strategy by considering the gain function $C_s(k) = \lambda_s \log_2(1 + \gamma_s(\mathbf{a}_k, \tau_k))$. Fig. 2.21 shows the average of the optimal performance obtained by considering the actual instantaneous gain at each point along the trajectory, and that obtained by considering the average gain for each point along the trajectory, i.e., using (2.64). We can see that these values grow linearly with the fraction of available channel capacity of the macro cell, λ_M, but the slope

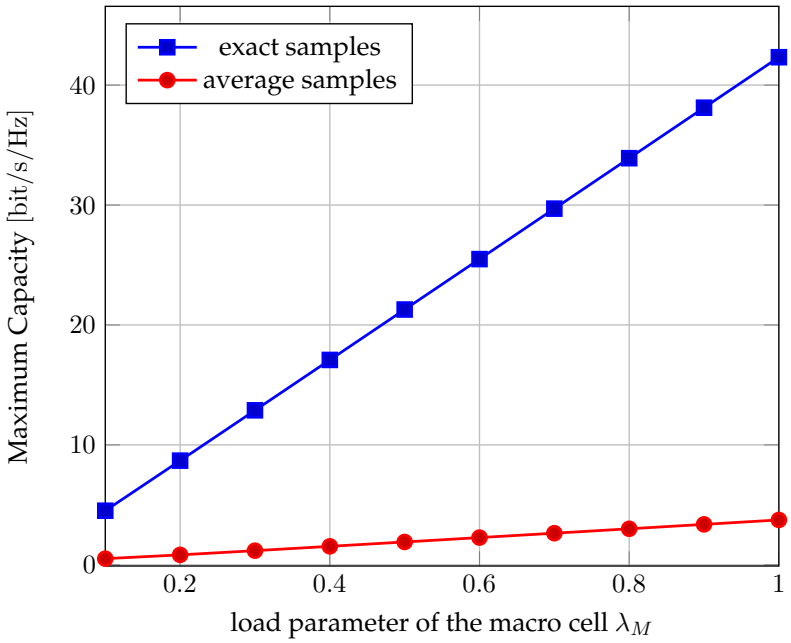

Figure 2.21. *Optimum capacity along the UE trajectory 2, with UE speed* $v = 40\ \mathrm{Km/h}$.

of the capacity curve computed from the exact samples is much higher than the other one. Hence, an accurate estimate of the fading conditions along the trajectory may potentially allow for significant performance improvements.

2.10 Handover Analysis in a multicell scenario

As a final remark on the HO analysis, we describe a possible extension of the mathematical model proposed in Sec. 2.5 to a scenario with multiple femtocells. We indicate with $\mathcal{F} = \{F_1, \ldots, F_N\}$ the set of N femtocells, placed within the macrocell coverage area. At every step of its trajectory, the UE can be connected either to one of the femtocells or to the macrocell, or can be switching from the serving to the target cell. The average capacity along the whole trajectory is still computed as in (2.36), except for the UE state space, which is now $\{M, H\} \cup \mathcal{F}$, i.e.,

$$\bar{\mathcal{C}}_L = \frac{2}{\pi} \int_0^{\pi/2} \frac{1}{N_L} \sum_{k=1}^{N_L} \sum_{S \in \{M,H\} \cup \mathcal{F}} \bar{C}_S(\mathbf{a}_k(\omega)) \mathrm{P}_S[\mathbf{a}_k(\omega)] \, \mathrm{d}\omega \,. \tag{2.68}$$

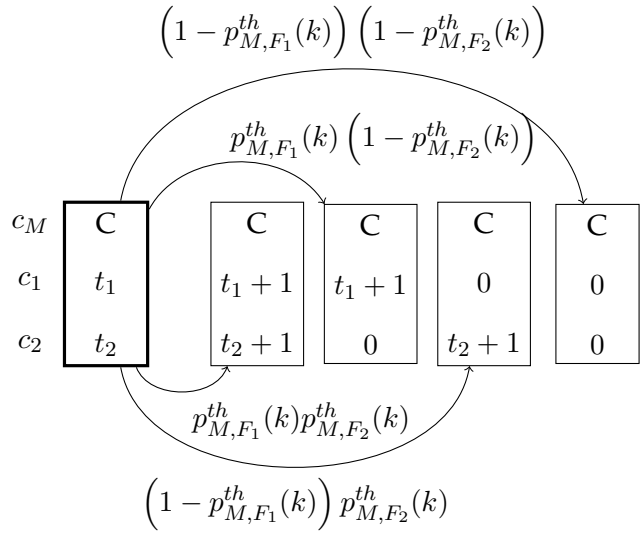

Figure 2.22. *Transitions from cell state* $< \mathrm{C}, t_1, t_2 >$ *(in bold), where* $0 \leq t_1, t_2 < N_T$.

The average capacity $\bar{C}_S(\mathbf{a}_k(\omega))$ at point $\mathbf{a}_k$ is given in (2.33) and (2.34), and the SINR $\gamma_S(\mathbf{a}_k, kT_c)$ with respect to the S-BS, $S \in \mathcal{F} \cup M$, is now given by

$$\gamma_S(\mathbf{a}_k, kT_c) = \frac{\Gamma_S(\mathbf{a}_k, kT_c)}{\sum_{S' \neq S} \Gamma_{S'}(\mathbf{a}_k, kT_c)}\,, \tag{2.69}$$

where each received signal has power as in (2.3).

The probability $\mathrm{P}_S[\mathbf{a}_k(\omega)]$ in (2.68) is defined as in Sec. 2.5 and computed from the Markov Chain described below.

The MC for the multi cell scenario is slightly more involved than the one for the single femtocell (see Fig. 2.7), but the principle of transition among states remains unchanged. The main difference is that we here need to take into account a TTT counter for each of the possible target BSs; the counter that expires first determines the next serving BS.

The states of the MC can be split into two classes. The first one describes the *cell states*, depicted with rectangular boxes in Fig. 2.22 and Fig. 2.23, where the UE is connected to any of the $N+1$ BSs and one or more TTTs can possibly start. We recall here that, according to the standard [7], the TTT from the UE serving cell $\mathcal{S}er$ towards the target cell $\mathcal{T}$ starts when the SINR

$$\gamma_{\mathcal{S}er,\mathcal{T}}(\mathbf{a}_k, kT_c) = \frac{\Gamma_{\mathcal{S}er}(\mathbf{a}_k, kT_c)}{\Gamma_{\mathcal{T}}(\mathbf{a}_k, kT_c)} \tag{2.70}$$

goes below threshold. In other words, in a multi-cell scenario the trigger condition involves the received powers of just the serving and the target BS. The cell states are defined as the

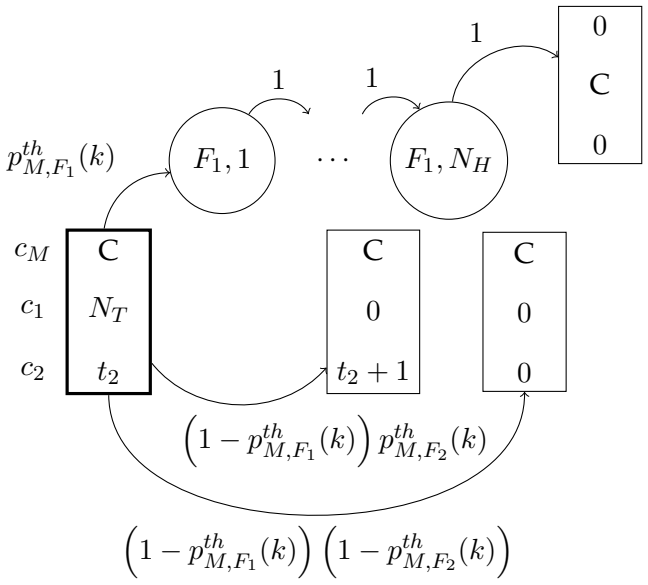

Figure 2.23. *Transitions from cell state* $< \mathrm{C}, t_1, t_2 >$ *(in bold), where* $t_1 = N_T$ *and* $0 \leq t_2 < N_T$.

$(N+1)$-tuples $< c_M, c_1, \ldots, c_N >$, where

$$c_S = \begin{cases} \mathrm{C} & \text{if } S = \mathcal{S}er \\ t & \text{otherwise}\,. \end{cases} \tag{2.71}$$

The parameter C indicates the BS that the UE is currently attached to, while the number $t \in \{0, 1, \ldots, N_T\}$ indicates for how many consecutive steps the SINR $\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c)$ has been below threshold, i.e., t represents the TTT counter for a possible handover to S-BS. The UE will be eventually connected to BS $S^* \neq \mathcal{S}er$ if $c_{S^*} = N_T$ and $\gamma_{\mathcal{S}er,S^*}(\mathbf{a}_k, kT_c)$ remains below threshold for one more step. Obviously, S^* is the state for which these conditions occur first.

The second class of states in the MC accounts for the handover procedures towards the new serving cell. In this case the *handover states*, depicted with circles in Fig. 2.23, are defined by the pair $< S, h >$ where S specifies the BS to be connected to and $h \in \{1, \ldots, N_H\}$ is the counter of the handover time.

For the sake of conciseness, we do not replicate here the rigorous analysis presented in Sec. 2.5 for the single cell case. We prefer instead to give some intuition on how the MC evolves in this more general case.

The transitions among cell states are constrained by the fact that, if at the k-th step $c_S = t$, with $t < N_T$ and $S \neq \mathcal{S}er$, then in the following step c_S could be either $t+1$, if

$\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c) < \gamma_{th}^{\mathcal{S}er}$, or 0 otherwise, i.e., the counter to S-BS is reset if its SINR goes above threshold. See Fig. 2.22 for an example of this transition in the case of $N = 2$ femtocells.

If instead $c_S = N_T$ and $\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c) < \gamma_{th}^{\mathcal{S}er}$, the UE starts the handover process to S-BS and the MC evolves to the handover state $< S, 1 >$. As before, the MC crosses deterministically all the handover states $< S, h >$, $h = 2, \ldots, N_H$, and ends up in the cell state where $c_S = \mathrm{C}$ and $c_{S'} = 0$, $\forall S' \neq S$. See Fig. 2.23 for an example of this transition in the case of $N = 2$ femtocells.

The probability $p_{\mathcal{S}er,S}^{th}(k)$ that the SINR $\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c)$ is below threshold is computed as in (2.40) and (2.41), and is equal to

$$p_{\mathcal{S}er,S}^{th}(k) = \mathrm{P}\left[\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c) < \gamma_{th}^{\mathcal{S}er}\right] = \frac{\gamma_{th}^{\mathcal{S}er}}{\gamma_{th}^{\mathcal{S}er} + \bar{\gamma}_{\mathcal{S}er,S}(\mathbf{a}_k)} \tag{2.72}$$

where

$$\bar{\gamma}_{\mathcal{S}er,S}(\mathbf{a}_k) = \frac{\Gamma_{\mathcal{S}er}^{tx}\, g_{\mathcal{S}er}(\mathbf{a}_k)}{\Gamma_S^{tx}\, g_S(\mathbf{a}_k)} \tag{2.73}$$

is the deterministic part of the SINR $\gamma_{\mathcal{S}er,S}(\mathbf{a}_k, kT_c)$.

Since the received powers from different cells are independent, the transition probabilities among the states of the MC are easily computed from (2.72) as the product of the probabilities with respect to all cells except the serving one, as can be seen from Fig. 2.22 and Fig. 2.23.

As a final comment, we note that the number of states N_{TOT} of the MC described above grows exponentially with the number of femtocells, since

$$N_{TOT} = (N + 1)(N_T^N + N_H)\,. \tag{2.74}$$

However, the complexity of the model can be reduced by considering only transitions among neighboring cells.

2.11 Summary

In this Chapter we showed the importance of a context-aware handover optimization in next generation cellular networks. We proposed two novel handover policies to maximize the user capacity along a random trajectory within a HetNet in different scenarios. The first provides the exact expression of the capacity with a simple channel propagation model,

while the second one computes the average capacity in a generic random channel environment. The latter exploits a novel analytical framework based on a Markov chain to consider the evolution of the UE state during the handover process and takes into account also the load condition of the cells. We showed that the performance obtained with the proposed policies outperforms a standard context agnostic handover policy. Finally, we derived and implemented an upper bound analysis to assess the residual margin of improvement of the proposed approaches with respect to the policy that provides the theoretically achievable maximum user performance.

Although in this thesis we assume that the UE trajectory is unknown, the proposed model can actually be adapted to account for exact (or statistical) knowledge of the UE path across the HetNet. In this case, the adoption of context-aware HO policies becomes even more crucial. The challenge, then, becomes the development of suitable techniques to estimate the context parameters, and the UE trajectory, in a simple and reliable manner, possibly using machine-learning approaches.

Chapter 3

Caching Strategies in Heterogeneous Networks

This chapter is organized as follows. Sec. 3.1 reviews our contribution compared to the state of the art on the proactive caching strategies. In Sec. 3.2 we present the detailed system model description, including the content request (Sec. 3.2.1), the content search (Sec. 3.2.2), and the user mobility (Sec. 3.2.3) processes. The average system cost is derived in Sec. 3.3, by computing the probabilities that a requested file is available either from another user through a D2D communication, or from the downlink with the closest small cell, or the macro cell, respectively. We express the system cost as a function of the caching variables, i.e., those variables that identify the content allocation at each user and BS cache. Moreover, the dependence on the user mobility pattern and the distribution of file interests are highlighted in the derived expression. The minimization of the average system cost is then formalized through a pseudoboolean optimization problem and solved with standard techniques, thus deriving the optimal caching strategy. However, due to the high complexity of this scheme, it is not possible to solve the problem for high values of the system parameters. In Sec. 3.4 we propose a more efficient yet suboptimal caching policy that can be used in more complex scenarios. Finally, in Sec. 3.5, we present numerical experiments to evaluate the performance of the optimal and suboptimal strategies, in comparison with a static caching strategy. Moreover, we study the impact of the context conditions, i.e., the user mobility probability, the skewness of content popularity, and the user interest profile on the system performance.

3.1 Related Work

The content placement problem within information centric networks (ICNs) has been widely explored and several solutions have been developed by optimizing different performance indicators, such as the hit ratio [52], the outage probability and the average delivery rate [53], the number of hops to deliver the requested content in the network, the link load condition, the social welfare [54], or the cost savings [55]. In [56], the authors propose a cost-aware caching strategy that aims at minimizing the operational costs needed by an Internet Service Provider (ISP) to retrieve the requested contents. The analysis developed in [52] takes into account the fact that content popularity can be dynamic over time, and computes the cache hit probability of standard caching policies, i.e., least recently used (LRU), q-LRU, and RANDOM, where the replacement of a newly arrived content is regulated by a deterministic, a semi probabilistic, and a pure probabilistic law, respectively. These analyses however do not exploit the potential of storing content at the small BSs which can provide higher cost benefits.

The paradigm of caching some popular files also at the small BSs has been developed to enhance the quality of service in a HetNet, and to alleviate the often congested links to the macro cell. The concept of femtocaching was introduced for the first time in [20], where user terminals can simultaneously access several small BSs, called helpers. The caching strategy, both coded and uncoded, is designed to minimize the expected download time of all files. The content placement framework at the small BSs is also studied in [23], where the content popularity profiles, assumed to be unknown, are estimated using the multi-armed bandit theory. In [57], a collaborative framework among small BSs is proposed, i.e., several files can be retrieved from the caches of different small BSs that belong to the same network domain. The addressed challenge is called in-network caching and consists in jointly studying the coalitions of small cells and the optimum file placement. The analysis in [24] considers the joint optimization of the content placement and routing problems, taking into account limited transmission capacity at the BSs.

The focus in these previous works is on caching strategies at the small cell, neglecting the possibility that also users can assist with their own caches. Furthermore, users are assumed to be static. Instead, we take user mobility into account and include cached files at the end users in our optimization framework.

Some recent papers study caching strategies in a dynamic scenario. The analysis in [58] extends [20] by introducing the concept of dynamic femtocaching, which consists in developing caching strategies in the presence of mobile users. The content allocation strategy developed in [59] exploits the user mobility information and assumes that each user can download parts of the requested content from different stations along its trajectory. In [60], instead, the authors focus on the user association problem, i.e., which mobile user should be connected to which BS at each time. The problem is solved as a one-to-many matching game, but according to the caching strategy the small BSs simply store the most popular files, which is suboptimal in our scenario. Moreover, no cache at the terminal side is assumed in the optimization problem.

Cache placement and cooperation techniques are studied in [61–64] by adopting a tree hierarchical cache topology. In [61] the authors focus on a 2-level hierarchical cache topology where users request some content that can be provided either by the leaf node they are connected to or by its parent node. If not available at any of the two locations, the file is delivered by the root node at a higher cost. In [62], the authors define two different types of cooperation, namely *intra-level* and *inter-level*, where files can be delivered only by other peer nodes, or only by parent nodes, respectively. The cache cooperation scheme in [63] extends this framework, allowing both types of cooperation simultaneously.

A promising approach is proposed in [65], where the authors assume that end users are partitioned in social wireless networks, within which they can exchange contents using a low cost message. When a file request is generated, the local cache is first checked and, if this search fails, the file is downloaded from the content provider's server using a standard 3G/4G cellular network. In that work user mobility is not taken into account in the optimization problem, leading to a static solution that is no longer valid in a dynamic scenario.

Strategies that entail caching at users and allow D2D communications have recently gathered a lot of interest [66]. The authors in [67] assume a static caching policy at the BSs, relays and users based on content popularity, and theoretically compute the average delivery rate and the outage probabilities of a typical user in different scenarios. Moreover, the network throughput is derived by assuming a maximum received-power cell association scheme. In [68], an optimal collaboration distance is defined, i.e., the spatial separation of two users that communicate through a D2D channel is theoretically derived in order to max-

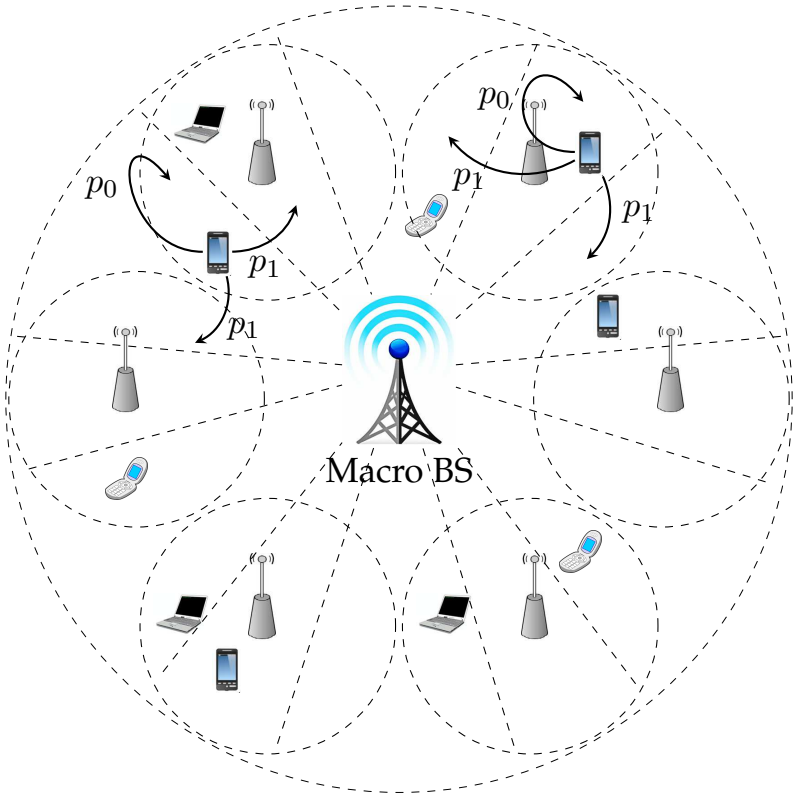

Figure 3.1. *Network scenario with a single macro cell, surrounded by $B = 6$ small cells deployed in a circle, and divided into $S = 3$ sectors. The arrows represent the possible user movement in one time slot, labeled with the respective probabilities, to remain in the current small cell (p_0) or to move to one of the adjacent small cells (p_1).*

imize the number of interference free links in a simple network with just one BS. This work has been extended in [69], where users are virtually grouped into clusters and share their cached files with other users in the same cluster. The macro BS manages the requests that can not be handled locally within clusters. In this thesis, we further extend the model in [69] by assuming also the presence of cache-provided small cells, deployed within the macro coverage area, and by developing a mathematical framework that keeps into consideration the user mobility.

3.2 System Model

The reference scenario is depicted in Fig. 3.1. There is one macro BS at the center of the network, which can communicate to all the users in the network. Small BSs are uni-

formly deployed in a circle around the macro BS, and each small BS is indicated as $b \in \mathcal{B} \equiv \{b_1, \ldots, b_B\}$. We assume that the coverage areas of the small BSs are non overlapping, thus a user can not be connected to multiple small cells simultaneously. Each small cell coverage area is partitioned into S sectors of the same area. In our model, a sector is indicated as $s \in \mathcal{S} \equiv \{s_1, \ldots, s_{B \cdot S}\}$. Each user $u \in \mathcal{U} \equiv \{u_1, \ldots, u_U\}$ is mobile and always connected to the macro BS, the closest small BS, b, and to all the users in the same sector s, where this set of users is named $\mathcal{U}^s$. In the following, we denote as ℓ_u the sector in which user u is located, and with b^s we identify the index of the small BS that includes sector s.

Time is divided into slots that are labeled with a discrete index $t \in \mathbb{N}$. Each slot is divided into a *user request phase* and a *cache replacement phase*, as in [23]. During the first phase, each user requests a file. If the file is available in the local cache of one of the other users in the same sector, the file can be transmitted via a D2D link, otherwise it is delivered through a standard cellular link by the small cell or the macro cell. We assume that the requested file is received with no errors within the same time slot it is requested. Moreover, as in [69], the macro BS controls the D2D links, and informs the right user to deliver the corresponding file to the user requesting it.

In the second phase, which is considered of negligible duration, the caches of both users and small cells are refreshed, i.e., updated with possibly new files, while some other files are discarded in order not to exceed the maximum capacity of each cache. We realistically assume that a user can cache (in the next time slot) only the files that are already present in its cache in the current slot, or can substitute one of them with the file that has just been received.

Finally, we indicate with $\mathcal{C}_u(t)$ and $\mathcal{C}_b(t)$ the sets of files cached at time slot t by user u and small BS b, respectively. $\mathcal{C}_{\mathcal{U}^s}(t)$ represents the set of files cached by all the users in $\mathcal{U}^s$ (users located in sector s) at time t and files in $\mathcal{C}_{\mathcal{U}^s \setminus u}(t)$ are cached by users in $\mathcal{U}^s$, but are not present in the cache of user u. Hence, $\mathcal{C}_{\mathcal{U}^s \setminus u}(t) \cap \mathcal{C}_u(t) = \emptyset$.

3.2.1 Content Request Generation Model

At the beginning of a time slot, each user requests a file $f \in \mathcal{F} = \{1, \ldots, F\}$, taken from a library of size F. All the files have the same size, as in [23].

We assume that each user belongs to a specific *class of interest* $k \in \mathcal{K} = \{1, \ldots, K\}$, which

determines a ranking order of the file popularities. This assumption reflects the fact that humans have different interests in real life. As a consequence, the popularity of a file f depends on the class k of the user requesting that file. For each class k, we assume that file popularities follow the Zipf distribution[1], which has been widely used in the literature to model content popularity distributions [71]. According to the Zipf law, the probability that a file is requested by a user of class k is given by

$$\mathrm{P}[f|k] = \frac{i(f,k)^{-\alpha}}{\sum_{j=1}^{F} j^{-\alpha}}\,, \tag{3.1}$$

where $i(f,k)$ is the rank of such file within class k, and $\alpha \geq 0$ is a fixed parameter that describes the skewness of file popularity. If $\alpha = 0$, all the files have the same popularity, while in the case of high values of α, there are only a few popular files, while the others have a very low probability to be requested.

We denote with k_u the class of user u, and with $r_u(t)$ the file requested by u at time t.

3.2.2 Content Search Model

Upon a content request by user u, different actions can be taken depending on where the file is stored, and consequently, different costs are encountered. The requested file $r_u(t)$ is first searched in the local cache $\mathcal{C}_u(t)$ of user u. If this search fails, the presence of the file is checked in the local caches of the users co-located with user u in sector s, i.e., in set $\mathcal{C}_{\mathcal{U}^s \setminus u}(t)$. If this second search also fails, the presence of the file is checked in $\mathcal{C}_{b^s}(t)$, the cache of the small BS b^s connected to user u. Finally, if all the previous searches failed, the file is downloaded from the macro BS cache that has all the files in the library. We depict in Fig. 3.2 an illustrative example of the content search model, where user u_1 is requesting a content at t. We assign to each of the previous actions a given cost to retrieve the requested file. This cost takes into account, e.g., the total downloading delay, the signaling overhead, the consumed bandwidth, or the battery usage at the client side, depending on the scenario of interest and on the considered application.

We then define the *user cost* $W_u(t)$ of user u requesting a file at time t according to the

[1] Any other distributions can be applied to our model [70].

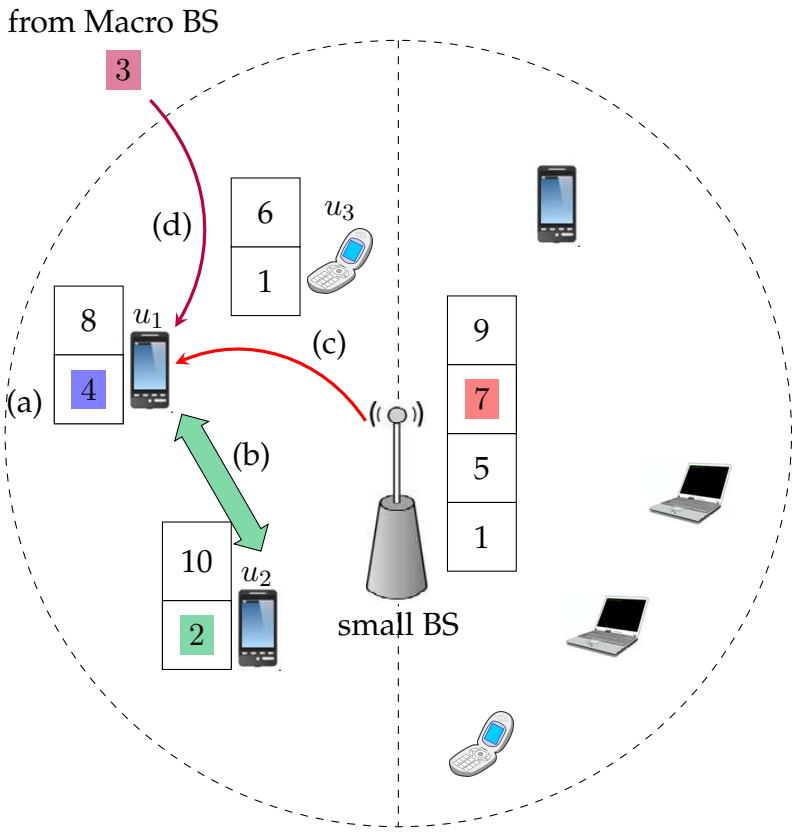

Figure 3.2. *Example of the content search model within a small cell with $S = 2$ sectors. If $r_{u_1}(t) = 4$ no action is taken (a); if $r_{u_1}(t) = 2$ a D2D communication (b) occurs between u_1 and u_2; if $r_{u_1}(t) = 7$, or $r_{u_1}(t) = 3$, the requested content is retrieved from the small BS (c), or the macro BS (d), respectively.*

cost of the action needed to deliver that file, i.e.,

$$W_u(t) = \begin{cases} w_0 & , \quad \text{if } r_u(t) \in \mathcal{C}_u(t) \\ w_1 & , \quad \text{if } r_u(t) \in \mathcal{C}_{\mathcal{U}^s \setminus u}(t) \\ w_2 & , \quad \text{if } r_u(t) \notin \mathcal{C}_{\mathcal{U}^s}(t) \wedge r_u(t) \in \mathcal{C}_{b^s}(t) \\ w_3 & , \quad \text{if } r_u(t) \notin \mathcal{C}_{\mathcal{U}^s}(t) \wedge r_u(t) \notin \mathcal{C}_{b^s}(t) \ , \end{cases} \tag{3.2}$$

where $w_0 \leq w_1 \leq w_2 \leq w_3$ are the increasing costs associated to each action.

3.2.3 User Mobility Model

The user mobility pattern is modeled as a discrete-time Markov model, where all the sectors are placed in a progressive sequence within the cell they belong to, such that each of them has exactly two neighboring sectors, as depicted in Fig. 3.1. E.g., the first sector of one cell has the last one of the previous cell and the second one of the same cell as neighboring sectors. We represent each sector as a state in a Markov chain. In each time slot a user can either stay in its current sector s, with probability p_0, move to the next sector, with probability p_1, or to the previous sector, with probability p_1, thus $p_0 + 2p_1 = 1$. We identify with $\mathbf{T}$ the $(B \cdot S) \times (B \cdot S)$ transition matrix of the Markov chain, which has the following structure

$$\mathbf{T} = \begin{bmatrix} p_0 & p_1 & 0 & 0 & \cdots & p_1 \\ p_1 & p_0 & p_1 & 0 & \cdots & 0 \\ 0 & p_1 & p_0 & p_1 & \cdots & 0 \\ 0 & 0 & p_1 & p_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ p_1 & 0 & 0 & \cdots & p_1 & p_0 \end{bmatrix} . \tag{3.3}$$

In this work we assume the same mobility pattern for all users. Different patterns can be taken into account by considering different transition matrices.

The adopted notation is summarized in Table 3.1.

3.3 Average System Cost

In this section we compute the system cost, which depends on the location, mobility, and content interests of each user, as well as on the caching strategy adopted.

Symbol	Definition
$\mathcal{U}$, U, u	User set, total number of users, and user index.
$\mathcal{B}$, B, b	Small BS set, total number of BSs, and BS index.
$\mathcal{S}$, S, s	Sector set, number of sectors per BS, and sector index.
$\mathcal{U}^s$	Set of users located in sector s.
b^s	Index of the small BS that includes sector s.
$\mathcal{C}_u(t)$, $\mathcal{C}_b(t)$	Set of files cached by user u and BS b at time t.
$\mathcal{C}_{b^s}(t)$	Set of files cached by the BS b^s at time t.
$\mathcal{C}_{\mathcal{U}^s}(t)$	Set of files cached by users in $\mathcal{U}^s$ at time t.
$\mathcal{C}_{\mathcal{U}^s \setminus u}(t)$	Set of files cached by users in $\mathcal{U}^s$ but not by u at time t.
$\mathcal{F}$, F, f	Set of all files, total number files, and file index.
$\mathcal{K}$, K, k	Set of classes of interest, total number of classes, and class index.
k_u	Interest class for u.
$\mathcal{U}_k$	Set of users with interest class k.
$i(f,k)$	Popularity rank of file f within class k.
α	Parameter of the Zipf distribution.
$r_u(t)$	File requested by u at time t.
$\ell_u\,(t)$	Sector where u is located at time t.
p_0 p_1	User probability to stay in the same sector, User probability to move to an adjacent sector.
$\mathbf{T}$	Transition matrix among sectors.
T_u^s	Probability that u moves to sector s in the next time slot.

Table 3.1. *Used notation.*

We define the system cost at time t as the sum of all the user costs:

$$\mathcal{W}(t) = \sum_{u \in \mathcal{U}} W_u(t) \ , \tag{3.4}$$

where the cost $W_u(t)$ for a single user was defined in (3.2).

At each time t, we assume to know the content of all caches, the location and the file requested by each user. With this information we can compute the system cost $\mathcal{W}(t)$. The goal is to find a caching strategy, for all users and small BSs, which minimizes the system cost $\mathcal{W}(t+1)$. The problem is that $\mathcal{W}(t+1)$ is expressed as a function of two independent random processes, since the mobility pattern and the requested files at $t+1$ are unknown. Thus, we choose to compute and minimize the expected system cost $\mathbb{E}\left[\mathcal{W}(t+1)\right]$. For the sake of notation, we skip in the following the dependence on the time index $t+1$, unless otherwise specified. The expected system cost is expressed as:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{W}\right] &= \sum_{u \in \mathcal{U}} \mathbb{E}\left[W_u\right] \\
&= \sum_{u \in \mathcal{U}} \Big\{ w_0 \mathrm{P}[r_u \in \mathcal{C}_u] + w_1 \mathrm{P}\left[r_u \in \mathcal{C}_{\mathcal{U}^s \setminus u}\right] \\
&\qquad + w_2 \mathrm{P}[r_u \notin \mathcal{C}_{\mathcal{U}^s} \wedge r_u \in \mathcal{C}_{b^s}] + w_3 \mathrm{P}[r_u \notin \mathcal{C}_{\mathcal{U}^s} \wedge r_u \notin \mathcal{C}_{b^s}] \Big\} \\
&= \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}} \mathrm{P}[\ell_u = s] \Big\{ w_0 \mathrm{P}[r_u \in \mathcal{C}_u] + w_1 \mathrm{P}\left[r_u \in \mathcal{C}_{\mathcal{U}^s \setminus u}\right] \\
&\qquad + w_2 \mathrm{P}[r_u \notin \mathcal{C}_{\mathcal{U}^s} \wedge r_u \in \mathcal{C}_{b^s}] + w_3 \mathrm{P}[r_u \notin \mathcal{C}_{\mathcal{U}^s} \wedge r_u \notin \mathcal{C}_{b^s}] \Big\} \\
&= \sum_{s \in \mathcal{S}} \sum_{f \in \mathcal{F}} \sum_{u \in \mathcal{U}} \mathrm{P}[\ell_u = s] \ \mathrm{P}[r_u = f] \Big\{ w_0 \mathrm{P}[f \in \mathcal{C}_u] + w_1 \mathrm{P}\left[f \in \mathcal{C}_{\mathcal{U}^s \setminus u}\right] \\
&\qquad + w_2 \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \in \mathcal{C}_{b^s}] + w_3 \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \notin \mathcal{C}_{b^s}] \Big\} \ ,
\end{aligned} \tag{3.5}$$

where in the last two equations we have conditioned on the location ℓ_u of user u and on its requested file r_u, respectively.

If we assume $w_0 = w_1$, i.e., if we do not distinguish the cost of having the file in user u's cache from a D2D communication cost, the last term in (3.5) can be simplified by writing

$$w_0 \mathrm{P}[f \in \mathcal{C}_u] + w_1 \mathrm{P}\left[f \in \mathcal{C}_{\mathcal{U}^s \setminus u}\right] = w_1 \mathrm{P}[f \in \mathcal{C}_{\mathcal{U}^s}] \ , \tag{3.6}$$

and it loses its dependency on u. Thus the system cost can be written as

$$\mathbb{E}\left[\mathcal{W}\right] = \sum_{s\in\mathcal{S}}\sum_{f\in\mathcal{F}} \Big\{ w_1 \mathrm{P}[f \in \mathcal{C}_{\mathcal{U}^s}] + w_2 \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \in \mathcal{C}_{b^s}] + w_3 \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \notin \mathcal{C}_{b^s}] \Big\} N_f^s \,, \tag{3.7}$$

where

$$N_f^s = \sum_{u\in\mathcal{U}} \mathrm{P}[\ell_u = s]\ \mathrm{P}[r_u = f] \tag{3.8}$$

$$= \sum_{u\in\mathcal{U}} T_u^s\, \mathrm{P}[f|k_u] \tag{3.9}$$

is the expected number of users within sector s (at time $t+1$) requesting file f. In (3.9), $T_u^s \triangleq \mathrm{P}[\ell_u = s]$ can be computed from the transition matrix in (3.3), since the location of u in the previous time slot is known, while $\mathrm{P}[f|k_u]$ is given in (3.1).

Letting $\mathbb{1}\{X\}$ be the indicator function of the event X, we denote with $\psi_u(f)$ and $\psi_b(f)$ the binary decision variables in the caching strategy, given by

$$\psi_u(f) = \mathbb{1}\{f \in \mathcal{C}_u\}\ , \forall u \in \mathcal{U} \tag{3.10}$$

$$\psi_b(f) = \mathbb{1}\{f \in \mathcal{C}_b\}\ , \forall b \in \mathcal{B}\,. \tag{3.11}$$

$\psi_u(f)$ and $\psi_b(f)$ are equal to 1 if the file f is present in the cache of user u and small BS b, respectively, and 0 otherwise.

We should now express the three probabilities in (3.7) as a function of $\psi_u(f)$ and $\psi_b(f)$, and then find the cache allocation strategy that minimizes the average system cost. The probability that file f belongs to the cache of at least one user located in sector s at time $t+1$ is computed in Appendix B.1, and is given by

$$\mathrm{P}[f \in \mathcal{C}_{\mathcal{U}^s}] = 1 - \prod_{u\in\mathcal{U}} \Big[1 - T_u^s\, \psi_u(f)\Big]\,. \tag{3.12}$$

The remaining two probabilities are easily derived from (3.12) as

$$\mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \in \mathcal{C}_{b^s}] = \mathbb{1}\{f \in \mathcal{C}_{b^s}\}\, \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s}] = \psi_{b^s}(f) \prod_{u\in\mathcal{U}} \Big[1 - T_u^s\, \psi_u(f)\Big]\,, \tag{3.13}$$

and

$$\mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s} \wedge f \notin \mathcal{C}_{b^s}] = \mathbb{1}\{f \notin \mathcal{C}_{b^s}\}\, \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s}] = [1 - \psi_{b^s}(f)] \prod_{u\in\mathcal{U}} \Big[1 - T_u^s\, \psi_u(f)\Big]\,. \tag{3.14}$$

Substituting (3.12)-(3.14) into (3.7), we can rewrite the system cost that can now be expressed in a more compact way as

$$\mathbb{E}\left[\mathcal{W}\right] = \overline{\mathcal{W}} + w \sum_{s\in\mathcal{S}} \sum_{f\in\mathcal{F}} N_f^s \left[1 - T_2\, \psi_{b^s}(f)\right] \prod_{u\in\mathcal{U}} \left[1 - T_u^s\, \psi_u(f)\right], \tag{3.15}$$

where

$$\overline{\mathcal{W}} = \sum_{s\in\mathcal{S}} \sum_{f\in\mathcal{F}} w_1 N_f^s \tag{3.16}$$

is the deterministic contribution of $\mathcal{W}$, while

$$w = w_3 - w_1\,, \quad T_2 = \frac{w_3 - w_2}{w_3 - w_1} \tag{3.17}$$

are positive constants that depend on the chosen weights.

We can formalize the proposed proactive caching policy in the form of a pseudoboolean optimization problem

$$\begin{aligned} \underset{\psi_u(f),\psi_b(f)}{\text{minimize}} \quad & \mathbb{E}\left[\mathcal{W}\right] \\ \text{subject to} \quad & \sum_{f\in\mathcal{F}} \psi_u(f) = |\mathcal{C}_u|\,,\ \forall u \in \mathcal{U} \\ & \sum_{f\in\mathcal{F}} \psi_b(f) = |\mathcal{C}_b|\,,\ \forall b \in \mathcal{B} \\ & \sum_{f\in\hat{\mathcal{F}}_u} \psi_u(f) = 0\,,\ \forall u \in \mathcal{U} \\ & \psi_u(f) \in \{0,1\}\,, \psi_b(f) \in \{0,1\}\,, \end{aligned} \tag{3.18}$$

where the first and second constraints of (3.18) guarantee that user's and small BS's caches do not exceed the maximum allowed capacity. In the third constraint, $\hat{\mathcal{F}}_u$ is the set of files that u can not cache at $t+1$ because it has not requested such files at $t+1$ and did not have them in its cache at t. Hence, the third constraint guarantees that files in $\hat{\mathcal{F}}_u$ are not available to be part of u's cache.

Unfortunately, problem (3.18) is highly non linear due to the several product terms within the cost (3.15) and not convex due to the binary constraints. A useful method (as in [72,73]) to transform (3.15) into a linear expression consists in substituting the product of the two binary variables ψ_{v_1} and ψ_{v_2} with a new binary variable $\psi' = \psi_{v_1}\psi_{v_2}$. If necessary, this procedure can be iterated multiple times till all the product terms are replaced with new single variables. To guarantee that the new variables are well defined, i.e., to force the

new variable to take the value of the product of the two substituted variables, the following additional inequality constraints are included in the problem (3.18): (i) $\psi' \geq 0$; (ii) $\psi' \leq \psi_{v_1}$; (iii) $\psi' \leq \psi_{v_2}$; (iv) $\psi' \geq \psi_{v_1} + \psi_{v_2} - 1$.

Indicating with $\boldsymbol{\psi}$ the vector of all the variables involved, including both the old and the new ones, problem (3.18) can be equivalently reformulated as the following linear problem

$$\begin{aligned} \underset{\boldsymbol{\psi}}{\text{minimize}} \quad & \mathbf{c}\boldsymbol{\psi} \\ \text{subject to} \quad & \mathbf{A}_{eq}\boldsymbol{\psi} = \mathbf{b}_{eq} \\ & \mathbf{A}\boldsymbol{\psi} \leq \mathbf{b} \\ & \psi_u(f) \in \{0,1\}, \psi_b(f) \in \{0,1\}. \end{aligned} \tag{3.19}$$

where $\mathbf{c}$ is the coefficient vector, while $\mathbf{A}_{eq}$, $\mathbf{A}$, $\mathbf{b}_{eq}$, and $\mathbf{b}$ are proper matrices and vectors used to represent the three equality constraints of (3.18), and the new inequality constraints.

The problem in (3.19) can be solved with standard integer programing optimization tools, even though, due to the enormous number of variables involved, a feasible solution can be found only for small values of U, S, B, and F.

3.4 Proactive Caching Policy

In this section, we introduce some simplifications that allow us to develop a suboptimal heuristic, with which we can efficiently determine the file allocation at the user and BS caches. Moreover, in Sec. 3.5 we will show that in a simple scenario the performance of our heuristic is almost equivalent to the optimal solution of the problem in (3.19).

First of all, we consider only one sector for each small cell, i.e., $S = 1$. In this way, we do not deal with the joint optimization among all the sectors that belong to the same BS. This is a particular case in which the coverage area of a small cell is small enough to allow D2D communications among all the users in the cell. The summation over the sector index s in (3.15) now becomes simply over the BS index b. With this simplification we can split the system cost into the contributions of several cell costs, as

$$\mathbb{E}[\mathcal{W}] = \sum_{b \in \mathcal{B}} \mathbb{E}\left[\mathcal{W}^b\right]. \tag{3.20}$$

The cell cost is given by

$$\mathbb{E}\left[\mathcal{W}^b\right] = \overline{\mathcal{W}^b} + w \sum_{f \in \mathcal{F}} N_f^b \left[1 - T_2\, \psi_b(f)\right] \prod_{u \in \mathcal{U}} \left[1 - T_u^b\, \psi_u(f)\right], \tag{3.21}$$

where

$$\overline{\mathcal{W}^b} \triangleq \sum_{f \in \mathcal{F}} w_1 N_f^b \tag{3.22}$$

does not influence the optimization problem and, consequently, our caching policy. In (3.21), T_u^b and N_f^b are the probability that user u moves to cell b and the expected number of users in cell b requesting file f, respectively.

3.4.1 BS cyclic optimization.

The idea at the basis of this suboptimal heuristic is to minimize separately the cell costs $\mathbb{E}\left[\mathcal{W}^b\right]$, $\forall b \in \mathcal{B}$. Since we consider, for every optimization problem referred to one cell, just one BS cache b and the few caches of the users that can move to b, the overall computational complexity is drastically reduced.

The issue with this approach is that the several subproblems, in which the main problem is divided, are not independent due to users' mobility. More precisely, since user u can reach multiple small cells (if $p_0 \neq 1$) in one time slot, the optimal solution for user u's cache should consider all its possible destinations. If we consider one BS at a time, we can obtain a different solution for u's cache for each BS.

To manage the possible incompatible file allocations for users' caches, we propose the following heuristic. We first find a partial solution to the problem by considering only the cache of one BS $\hat{b}$ and the caches of all the users that can possibly move to $\hat{b}$. As a remark, we reduce (3.21) to a linear expression as explained in the previous section, and we select from (3.19) only the constraints referred to the caches of $\hat{b}$ and of the users that can move to $\hat{b}$ in one time slot, and neglect the others. Then we proceed by considering the neighboring BS of $\hat{b}$ in counterclockwise order, and optimizing the caches of that small BS and of the users that can move to the corresponding cell. The users with an assigned content set can not change their cache anymore. We repeat the same rationale for all the BSs in the system in a cyclic order. Finally, we repeat the same steps above by starting the procedure each time from a different BS. We eventually select the content allocation that provides the minimum system cost in (3.15).

3.5 Performance Evaluation

In this section, we simulate the network scenario to evaluate the performance of the proposed proactive policies, i.e., the optimal solution of the problem (3.19), and the heuristic strategy described in Sec. 3.4. We assume that users are initially distributed uniformly at random among the different small cells. Each user can store exactly one file, i.e., $|\mathcal{C}_u| = 1$, $\forall u \in \mathcal{U}$, in accordance to the work in [68, 69]. For each user u at the end of each time slot $t+1$, the caching policy should decide whether to keep the file that was stored in u's cache at time t, or the one that has been requested and downloaded by user u at $t+1$, while all the other files are not available to be stored.

The capacity of the cache of each small BS is instead $|\mathcal{C}_b| = 5\,, \forall b \in \mathcal{B}$. The number of user's classes of interest is $K = 2$, while the ranking order within a single class is chosen uniformly among all the permutations of file popularities. The cost vector is chosen to strongly penalize the request of files to the macro BS, i.e., $[w_1\,, w_2\,, w_3] = [1\,, 10\,, 100]$. However, we stress the fact that the proposed policies can work with an arbitrary cost vector.

In order to evaluate the performance of the proposed proactive policies, we introduce a *static policy* that serves as a comparison. The static policy keeps in the cache of each small BS the $|\mathcal{C}_b|$ most globally popular files, considering the population of all the users, according to the aggregated rank

$$I(f) = \sum_{k=1}^{K} |\mathcal{U}_k|\, i(f,k)\,, \tag{3.23}$$

where $\mathcal{U}_k$ is the set of users with interest class k, and $i(f,k)$ is the rank of file f for that class. In the cache of each user the static policy stores $|\mathcal{C}_u|$ random files, chosen according to the Zipf distribution of u's interest.

We compute the system cost $\mathcal{W}_{pro}(t)$ of both the proactive policies and the one $\mathcal{W}_{sta}(t)$ of the static policy, during a time window of $t_{max} = 20$ time slots. We define the gain G of a proactive policy as the sum of the difference between the proactive and the static costs, relative to the integral of the static cost, as given by

$$\mathrm{G} = \frac{\sum_{t=1}^{t_{max}} \left[\mathcal{W}_{sta}(t) - \mathcal{W}_{pro}(t)\right]}{\sum_{t=1}^{t_{max}} \mathcal{W}_{sta}(t)}\,. \tag{3.24}$$

Due to the high complexity of the optimum algorithm, we consider a scenario with $B = 10$ small BSs, $U = 30$ users and a library of $F = 500$ files. In Fig. 3.3, we plot the average

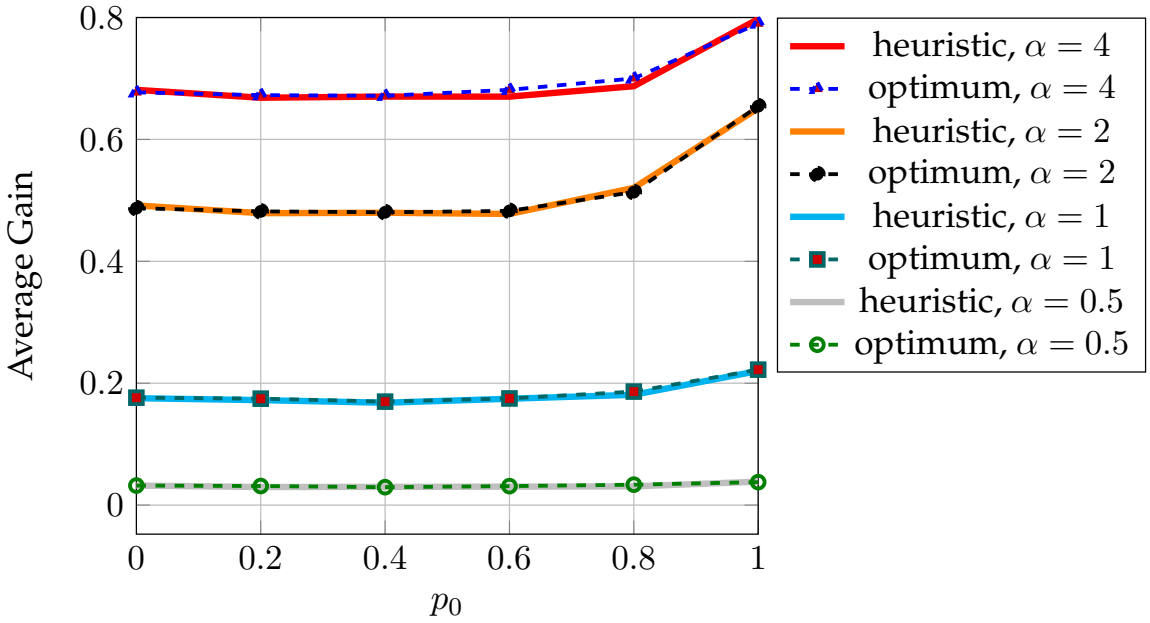

Figure 3.3. *Average gain of the optimum proactive caching policy and the proposed heuristic with respect to the static policy, as a function of the probability that a user does not change location in the next time slot. System parameters are* $B = 10$, $U = 30$, *and* $F = 500$.

gain (3.24), as a function of the probability that each user does not change location in the next time slot (p_0), for $\alpha = \{0.5, 1, 2, 4\}$. The values of G are averaged over 1000 iterations.

By looking at the results, we observe that the performance of the heuristic is almost equivalent to the optimal one, for all the values of α. If compared to the static policy, the average gain of the proactive strategies is constant for values of $p_0 < 1$, while it reaches the maximum when $p_0 \simeq 1$. This is the case of reduced mobility, where the probability of changing small cell in the next time slot is very small. In this case, the cell cost is not affected by the random mobility pattern.

We finally compute the fraction of files downloaded with a D2D communication, from the small BS, or from the macro BS, and we indicate these three cases with

$$e \in \{\text{“D2D”}, \text{“small BS”}, \text{“macro BS”}\}, \tag{3.25}$$

respectively. The fraction corresponding to each case e is given by

$$\mathrm{N}_e = \frac{1}{t_{max} U} \sum_{t=1}^{t_{max}} \sum_{u \in \mathcal{U}} \mathbb{1}\{r_u(t) \text{ is provided by } e\} . \tag{3.26}$$

In Fig. 3.4 and Fig. 3.5 we plot the fractions from (3.26) with $p_0 = 0.6$ and $p_0 = 1$, and

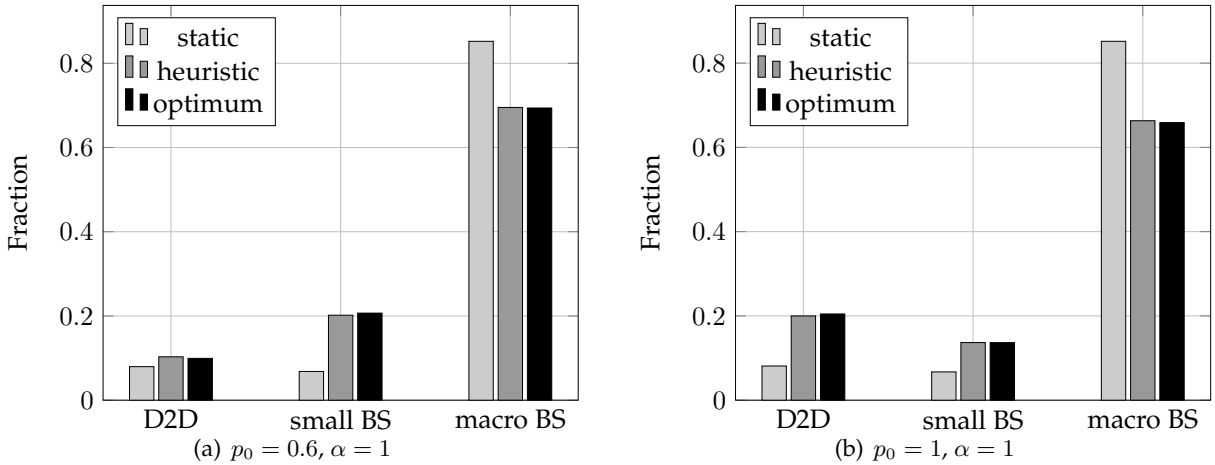

Figure 3.4. *Fraction of D2D, small BS and macro BS transmissions for the static, the optimal proactive and the heuristic caching policies. System parameters are $B = 10$, $U = 30$, and $F = 500$.*

$\alpha = 1$, $\alpha = 2$, respectively, in the same scenario of above. We average the value of N_e over 1000 iterations.

From Fig. 3.4(a), we observe that the gain achieved by the proactive policies for $\alpha = 1$ is mainly due to the smart file allocation at the small BS caches, while most of these files need to be requested to the macro cell in the case of the static policy. In other words, the proactive policies tend to be conservative and place the most requested files at the small BSs, rather than at the user caches. This is mainly due to user mobility that does not guarantee that a popular file at the user cache can be used also by another user if they end up in two separate sectors. It is hence preferred to keep the most common files at the small BS's cache. The problem with the static policy is that the most common files are placed both at the users' and at the BSs' caches, without coordination and with the drawback of file duplication. Hence, a smaller number of files is available locally, and some files are inevitably requested at the macro cell, increasing the system cost.

A different configuration is given in Fig. 3.4(b), where D2D communications are more frequent for our policies with respect to the static one. In this case in fact users do not move and the cluster they form within a small cell is fixed for every time slot, encouraging the D2D activations whenever possible. The most requested files are hence placed at the users cache, while the small cell caches are subsequently filled with the less popular files, avoiding duplication with the content at the user caches. Since the cost of delivering contents from the small BS is higher than the one of a D2D communication, the average gain for $p_0 = 1$ is

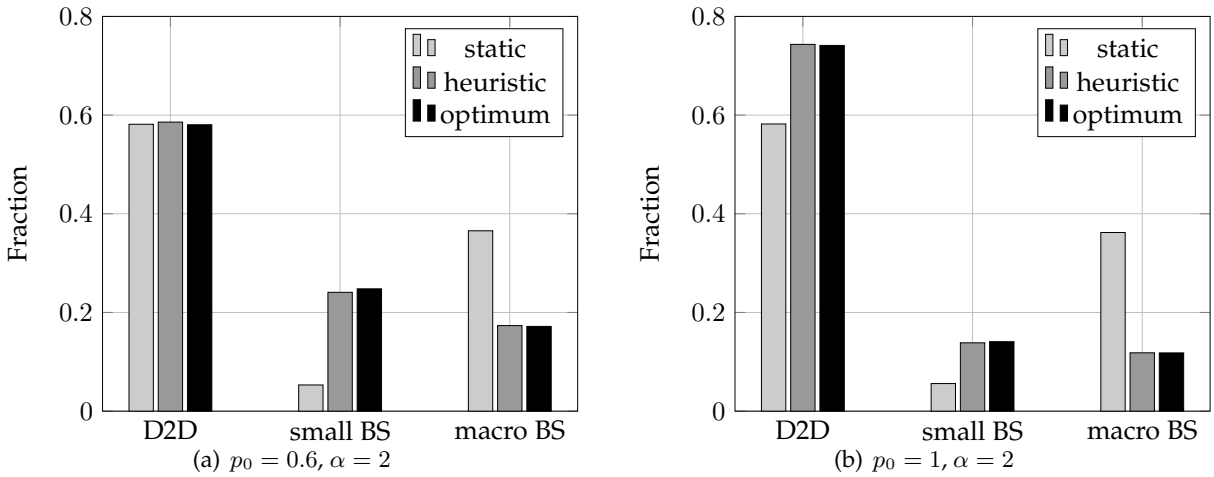

Figure 3.5. *Fraction of D2D, small BS and macro BS transmissions for the static, the optimal proactive and the heuristic caching policies. System parameters are* $B = 10$, $U = 30$, *and* $F = 500$.

more significant than the case with $p_0 = 0.6$, as can be seen from Fig. 3.3.

For the sake of comparison, we plot in Fig. 3.5 the fractions obtained in the same scenario of Fig. 3.4 but with $\alpha = 2$. In this case the interests of the users are such that only a limited number of files is requested with a very high probability, while most of the files are only rarely requested. Thus, the requests to the macro BS halved for the static policy as compared to the case in which $\alpha = 1$. For the heuristic and the optimum policy, this predictable request pattern is even more beneficial, allowing them to minimize the requests to the macro BS, whose fraction falls below 0.2.

Finally, in Fig. 3.6, we plot the average gain of our heuristic when varying the parameter α of the Zipf distribution. The gain is zero in the two extreme situations, i.e., if the contents are requested with a uniform probability ($\alpha = 0$) and if instead the most popular file alone has more than 99% of probability to be requested ($\alpha > 10$). In the first case, in fact, our policy performs poorly due to the high level of uncertainty associated with the content request process, while in the second case the static policy reaches the performance of our proactive strategy since choosing the most popular file at the users' caches is the optimal strategy almost always.

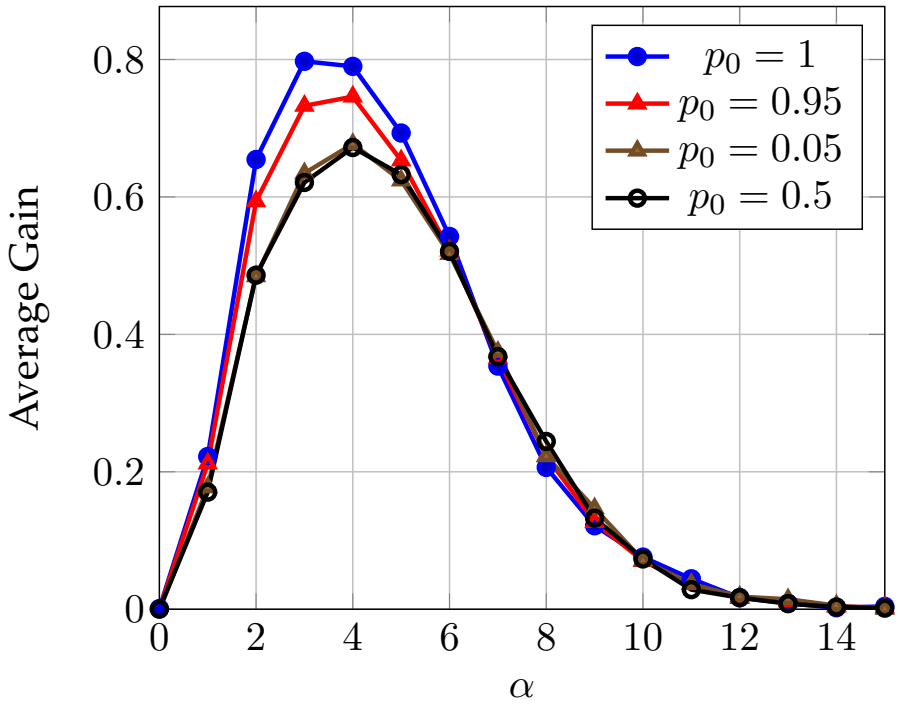

Figure 3.6. *Average gain of the proposed heuristic with respect to the static policy, as a function of the Zipf distribution parameter α. System parameters are $B = 10$, $U = 30$, and $F = 500$.*

3.6 Summary

In this Chapter we designed a proactive caching policy for a HetNet scenario to jointly optimize the choice of files to be stored at the mobile users and at the small BSs. We derived a closed form expression for the average system cost as a function of the user mobility level within HetNets and of the content request distribution. Then, we proposed two caching policies. The first leads to an optimal solution that minimizes the overall system cost but is computationally infeasible for large numbers of users and BSs. The second one is a heuristic that leads to a suboptimal solution and requires less computation resources. We proved that the heuristic indeed performs almost as well as the optimal policy and showed that both policies outperform a static reactive policy that does not take the context information into account.

Several aspects are currently considered for further investigation. First of all, we plan to extend the proposed policy in the case of an arbitrary number of sectors within the same small cell, leading then to a more generic solution. Moreover, the influence of other context parameters, as the number of classes of interests, the file ranking order, and the chosen

weights related to different delivery modes, are useful to derive comprehensive proactive caching strategies.

Chapter 4

Conclusions

In this thesis a set of optimization algorithms were designed to manage the handover mechanism and the caching policy within a HetNet. We adopted several analytical and simulation tools to provide original solutions for resource management problems. Moreover, we investigated the impact of the context information on the overall performance, thus developing context-aware optimization.

In Chapter 2 of this thesis we proposed a novel approach to optimize the handover procedure in HetNets by considering context parameters, such as the user speed, the channel gains and the load information of the cells. We derived two novel analytical frameworks to compute the average Shannon capacity along the UE trajectory. The first one assumes a simple propagation channel model and derives the exact trajectory average capacity. The second one instead assumes a general transmission channel and makes use of a Markov chain to model the evolution of the UE state during the handover process. The models were then used to derive our handover strategy, namely CAHP, that maximizes the UE average capacity in different scenarios, as a function of the context parameters. By adding suitable offsets to the HO thresholds, we then adjusted the mathematical model and the CAHP algorithm to account for the traffic loads of the cells. We presented a number of simulation results to assess the performance obtained by the proposed policy in comparison with standard HO policies with fixed TTT.

We then proposed a simple but effective mathematical framework to assess the theoretical optimal HO performance in a given context. The model has been used to derive the optimal performance in a sample scenario, thus providing a benchmark to assess the perfor-

mance of some practical algorithms taken from the literature, including the proposed CAHP. As a final result, we show that our model can be easily adapted to derive the HO analysis in a multicell scenario.

From our study it clearly emerges that context-awareness can indeed improve the handover process and significantly increase the performance of mobile UEs in HetNets.

In Chapter 3 of this thesis, we studied proactive content allocation strategies for a HetNet environment, where mobile users and small BSs are provided with storage capabilities. The proposed solution exploits the possibility of content delivery either through device-to-device communication among users or through cellular links from the nearby BSs. Firstly, we derived the closed form expressions of the probability that the requested content is delivered by a user, a small cell, or a macro cell. Then, we analytically derive the average system cost, based on a set of context parameters that describe both the user mobility and the content request process. Moreover, the penalty associated to each delivery procedure can be arbitrarily regulated, thus allowing to accommodate any objective metric.

We have developed two algorithms for the optimal cache allocation that jointly optimize the content placement at the users and BSs caches. More precisely, we have proposed an optimal policy, that minimizes the overall average system cost and can be used in scenarios with limited number of users and BSs, and a heuristic, that combines the optimal allocations derived separately at each small BSs. Even if the heuristic is suboptimal, we have shown that it performs almost identically to the optimal policy and hence can be used to derive a proactive caching policy in complex scenarios, with a large number of users, BSs and files. We compared the performance of the proposed strategies with a static reactive caching policy that keeps the cache allocation unchanged through time. The significant gain we obtained revealed the importance of context information within the optimization framework. We finally studied the impact of mobility and content request distribution on the optimal cache allocation and as a consequence on the final performance.

Appendix A

Appendix related to Chapter 2

A.1 Computation of the internal trajectory capacity in Sec. 2.3.1.1

From the intervals $I_n(a,S) = [\alpha_n(a,S), \beta_n(a,S)]$ specified in Table 2.1, (2.19) can be computed as

$$C_{L,int} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^R C_S(a) \sum_{n=0}^{2} n \int_{I_n(a;S)} \frac{1}{\sqrt{1-(R/a)^2\sin^2\omega}}\,\mathrm{d}\omega\,\mathrm{d}a \tag{A.1}$$

$$= \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^R C_S(a) \sum_{n=0}^{2} n \left[F\left(\beta_n(a,S), \frac{R}{a}\right) - F\left(\alpha_n(a,S), \frac{R}{a}\right)\right] \mathrm{d}a\,, \tag{A.2}$$

where $F(\phi,k)$ is the *incomplete elliptic integral of the first kind*, defined as

$$F(\phi,k) = \int_0^\phi \frac{1}{\sqrt{1-k^2\sin^2\omega}}\,\mathrm{d}\omega\,. \tag{A.3}$$

After simple algebra, since $\alpha_2(a,S) \equiv \beta_1(a,S)$ and $\alpha_1(a,S) = 0$, we obtain

$$C_{L,int} = \sum_{S\in\{M,F,H\}} \frac{2}{L\pi} \int_0^R C_S(a) \left[G\left(\beta_1(a,S), \beta_2(a,S), \frac{R}{a}\right) - F\left(\beta_2(a,S), \frac{R}{a}\right)\right] \mathrm{d}a\,, \tag{A.4}$$

where $G(\phi_1,\phi_2,k) = F(\phi_2,k) - F(\phi_1,k)$.

A.2 Closed form expression of the average capacity (2.33)

From (2.39), the probability density function of ξ is given by

$$f_\xi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{P}[\xi \le x] = \frac{1}{(x+1)^2}\,, \quad x \in [0,+\infty]\,. \tag{A.5}$$

Given $\bar{\gamma}$, the expectation of $\log_2(1+\bar{\gamma}\xi)$ is computed as

$$\begin{aligned}\int_0^{+\infty} \log_2\left(1+\bar{\gamma}x\right) f_\xi(x)\,\mathrm{d}x &= \int_0^{+\infty} \log_2\left(1+\bar{\gamma}x\right) \frac{1}{(x+1)^2}\,\mathrm{d}x \\ &= -\beta\,\frac{\ln\left(1+\bar{\gamma}x\right)}{1+x}\bigg|_0^{+\infty} + \beta\bar{\gamma}\int_0^{+\infty} \frac{1}{1+x}\,\frac{1}{1+\bar{\gamma}x}\,\mathrm{d}x \\ &= \beta\frac{\bar{\gamma}}{\bar{\gamma}-1}\int_0^{+\infty}\left[\frac{\bar{\gamma}}{1+\bar{\gamma}x} - \frac{1}{1+x}\right]\mathrm{d}x \\ &= \frac{\bar{\gamma}}{\bar{\gamma}-1}\log_2\left(\frac{1+\bar{\gamma}x}{1+x}\right)\bigg|_0^{+\infty} \\ &= \frac{\bar{\gamma}}{\bar{\gamma}-1}\log_2\left(\bar{\gamma}\right)\end{aligned}$$

where $\beta = \log_2 e$ and integration by parts was used to solve the integral.

A.3 Closed form expression of the average capacity in (2.64)

In the following we derive expressions (2.65) and (2.66). For the sake of simplicity, we omit the dependence on $\mathbf{a}_k$ and τ_k. The cumulative distribution function (CDF) of γ_s is computed as

$$\begin{aligned}F_{\gamma_s}(x) &= \Pr\left[\gamma_s \le x\right] \\ &= \Pr\left[\alpha_s \le \frac{x}{\bar{\Gamma}_s}\left(\sum_{i\in\mathcal{B}\setminus s}\bar{\Gamma}_i\alpha_i\right)\right] \\ &= 1 - \prod_{i\in\mathcal{B}\setminus s}\int_0^{+\infty} f_{\alpha_i}(y_i)e^{-xy_i\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}\,\mathrm{d}y_i \\ &= 1 - \prod_{i\in\mathcal{B}\setminus s}\int_0^{+\infty} e^{-\left(1+x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)y_i}\,\mathrm{d}y_i \\ &= 1 - \prod_{i\in\mathcal{B}\setminus s}\frac{1}{1+x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}\,.\end{aligned}$$

The probability density function (PDF) of γ_s is given by

$$\begin{aligned} f_{\gamma_s}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\gamma_s}(x) \\ &= \frac{\sum_{i\in\mathcal{B}\setminus s} \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s} \prod_{j\in\mathcal{B}\setminus\{s,i\}} \left(1 + x\frac{\bar{\Gamma}_j}{\bar{\Gamma}_s}\right)}{\prod_{i\in\mathcal{B}\setminus s} \left(1 + x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)^2} \\ &= \sum_{i\in\mathcal{B}\setminus s} \left(\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)^N \frac{1}{\prod_{j\in\mathcal{B}\setminus\{s,i\}} \left(\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s} - \frac{\bar{\Gamma}_j}{\bar{\Gamma}_s}\right)} \frac{1}{\left(1 + x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)^2} \\ &= \sum_{i\in\mathcal{B}\setminus s} \underbrace{\frac{1}{\prod_{j\in\mathcal{B}\setminus\{s,i\}} \left(1 - \frac{\bar{\Gamma}_j}{\bar{\Gamma}_i}\right)}}_{\psi_{s,i}} \frac{\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}{\left(1 + x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)^2}. \end{aligned}$$

Finally, the expectation in (2.64) is computed as

$$\begin{aligned} \mathbb{E}\left[\log_2\left(1+\gamma_s\right)\right] &= \int_0^{+\infty} f_{\gamma_s}(x) \log_2(1+x)\, \mathrm{d}x \\ &= \log_2 e \int_0^{+\infty} \sum_{i\in\mathcal{B}\setminus s} \psi_{s,i} \frac{\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}{\left(1 + x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}\right)^2} \ln(1+x)\, \mathrm{d}x \\ &= \log_2 e \sum_{i\in\mathcal{B}\setminus s} \psi_{s,i} \left[-\frac{\ln(1+x)}{1 + \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}x}\Big|_0^{+\infty} + \int_0^{+\infty} \frac{1}{1 + \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}x} \frac{1}{1+x}\, dx \right] \\ &= \log_2 e \sum_{i\in\mathcal{B}\setminus s} \psi_{s,i} \int_0^{+\infty} \left[\frac{\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}{1 + \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}x} - \frac{1}{1+x} \right] dx \\ &= \log_2 e \sum_{i\in\mathcal{B}\setminus s} \frac{\psi_{s,i}}{1 - \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}} \ln\left(\frac{1+x}{1 + x\frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}}\right)\Bigg|_0^{+\infty} \\ &= \sum_{i\in\mathcal{B}\setminus s} \frac{\psi_{s,i}}{1 - \frac{\bar{\Gamma}_i}{\bar{\Gamma}_s}} \log_2 \frac{\bar{\Gamma}_s}{\bar{\Gamma}_i}. \end{aligned}$$

Appendix B

Appendix related to Chapter 3

B.1 Probability (3.12) that file f is part of the cluster of users within sector s.

We report here the computation of (3.12) of the probability that file f is part of the cluster of users within sector s.

$$\begin{aligned}
\mathrm{P}[f \in \mathcal{C}_{\mathcal{U}^s}] &= 1 - \mathrm{P}[f \notin \mathcal{C}_{\mathcal{U}^s}] && \text{(B.1)}\\
&= 1 - \prod_{\substack{u \in \mathcal{U} \text{ s.t.} \\ T_u^s \neq 0}} \mathrm{P}[f \notin \mathcal{C}_u] && \text{(B.2)}\\
&= 1 - \prod_{u \in \mathcal{U}} \left\{ \mathrm{P}[f \notin \mathcal{C}_u | \ell_u = s] \, \mathrm{P}[\ell_u = s] + \underbrace{\mathrm{P}[f \notin \mathcal{C}_u | \ell_u \neq s]}_{=1} \mathrm{P}[\ell_u \neq s] \right\} && \text{(B.3)}\\
&= 1 - \prod_{u \in \mathcal{U}} \left[(1 - \mathbb{1}\{f \in \mathcal{C}_u\}) \, T_u^s + 1 - T_u^s \right] && \text{(B.4)}\\
&= 1 - \prod_{u \in \mathcal{U}} \left[1 - T_u^s \, \psi_u(f) \right], && \text{(B.5)}
\end{aligned}$$

where in (B.3) we use the total probability theorem, while in (B.4) we notice that probability $\mathrm{P}[f \notin \mathcal{C}_u | \ell_u = s]$ reduces to the binary variable $\mathbb{1}\{f \notin \mathcal{C}_u\}$, which depends on the caching policy that can assign file f to the cache of user u. This allows us to distinguish the policy-dependent term $\psi_u(f)$, that we have to optimize, from the mobility probability T_u^s, that depends on the model, instead.

Appendix C

Bayesian Machine Learning Inference of Video Dynamic Characteristic

In this Appendix, we consider the problem of the inference of video dynamicity according to a Bayesian machine learning approach. The knowledge of the dynamic characteristic of a certain video, i.e., whether the video scenes are fast moving rather than static, is essential for the video resource allocation, e.g., the transmission rate to be used to transmit the video over the Internet. The increasing of data traffic, especially of video content, has raised the need of an effective radio resource usage and optimization. Motivated by the fact that existing resource allocation algorithms do not consider the video dynamic characteristics, i.e., they are content-agnostic, we aim at developing a resource allocator that is content-aware instead. Hence, we build a model that firstly predicts the video motion level and then exploiting this information selects a proper transmission rate of the video.

In this Appendix, we present our preliminary work on the prediction of the video dynamic characteristic. In Sec. C.1 we introduce some basic concepts that describe the structure of a coded video sequence. In Sec. C.2 we build a Bayesian Network framework to infer the dynamicity level of a certain video from its frame sizes, that are the only observable variables in our model. More specifically, we distinguish three different levels of video motion and use a naive Bayesian classifier for each of them. The prediction of the hyperparameters of the Dirichlet distribution used for the prior is discussed in Sec. C.3, while some simulation results are derived in Sec. C.4.

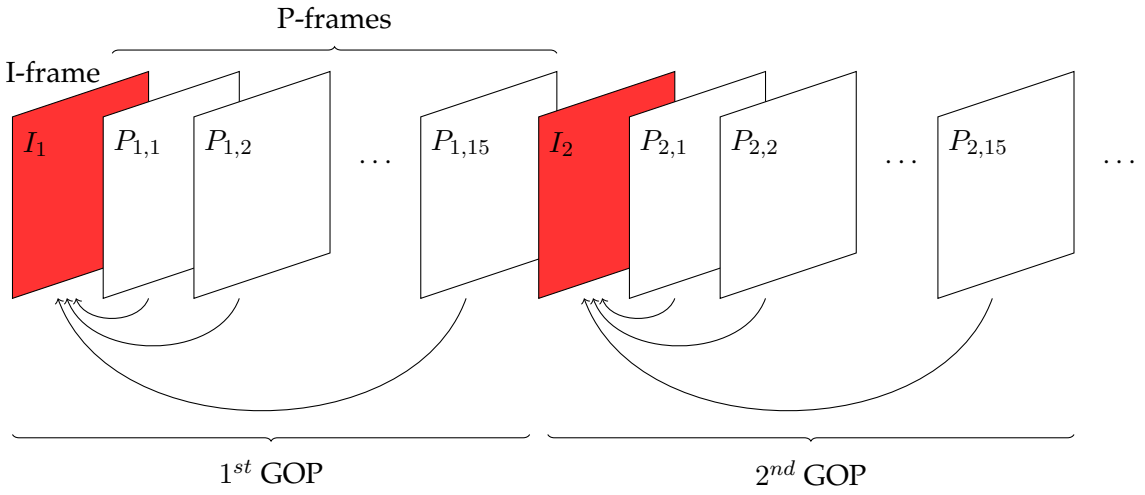

Figure C.1. *GOPs structure in a coded video sequence.*

C.1 Introduction

Typically, the compression of a video stream is realized by dividing its whole frame sequence into groups of 16 frames, where each group is encoded independently of the others. A group of 16 encoded frames is called Group-Of-Pictures (GOP) and has a fixed structure. As shown in Fig. C.1, we consider a simple encoding scheme where the first frame of a GOP, called Intra coded frame (I-frame), is obtained through the standard JPEG compression of the original raw frame, while the following 15 frames of the same GOP, called Predicted coded frames (P-frames), are coded through their difference from the corresponding I-frame. Letting N_G be the total number of GOPs in the video sequence, we indicate with I_m and $P_{m,n}$ the I-frame within the m-th GOP and the n-th P-frame in the same GOP, respectively, where $m \in \{1, 2, \ldots, N_G\}$ and $n \in \{1, 2, \ldots, 15\}$. In the following we will use symbols I_m and $P_{m,n}$ to indicate both video frames and their sizes.

Since the encoding algorithm includes both frame image compression and motion compensation techniques, the distribution of the frame sizes intrinsically carries information about the complexity and dynamicity of the video. Generally, I-frames have higher sizes than the corresponding P-frames which capture just the differences in the video motion and are coded with a higher level of compression. However, if the video scenes are fast-moving, consecutive frames would be very different from each other, and especially the P-frames of a GOP would differ a lot from the corresponding I-frame, resulting thus in high sizes. On the other hand, static videos have similar consecutive frames and, as a consequence, their

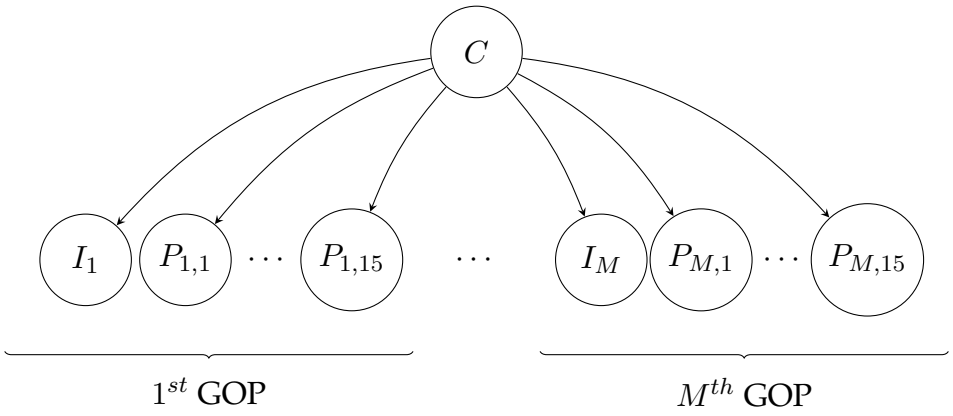

Figure C.2. *Naive Bayesian Network of our model.*

P-frames have on average lower sizes compared to dynamic videos because subjected to a higher compression. From the rationale above, it is evident that the frame sizes depend on the dynamic characteristic of the video and that the relationship between these parameters is too complex to be evaluated using common analytical tools.

C.2 System Model

We propose a system model based on a naive Bayesian classifier as in Fig. C.2 to infer the dynamicity Class C given the frames size of some GOPs of the video. In this preliminary work we assume that the frame sizes are conditionally independent given the class, leaving a structure learning analysis as a future investigation. According to [74], and as proved in [75] this kind of structure is the proper choice when the sample data size is small, as in our case. A more accurate model would require more parameters and, as a consequence, would increase the sensitivity of the estimation variance, which increases as the sample size decreases.

In Fig. C.2, $C \in \{1, 2, 3\}$ is the random variable that represents the video dynamics. For simplicity, we consider just three levels of motion, i.e., static ($C = 1$), medium ($C = 2$), and dynamic ($C = 3$). The video dynamic characteristic is retrieved from previous works on video classification using the Structural Similarity (SSIM) indicator curves and clustering techniques. I_m and $P_{m,n}$ denote the frame sizes where $m \in \{1, 2, \ldots, M\}$, $n \in \{1, 2, \ldots, 15\}$, and M is the maximum index among the observable GOPs, $M \leq N_G$. More precisely, I_m and $P_{m,n}$ are the discretized sizes obtained after quantization of their real values. Since I-frames

have on average a higher size distribution than P-frames, we use two different quantizers that have different dynamic ranges but equal number of quantization intervals L for the two types of frames. In particular $[4, 12] \times 10^4$ bytes and $[4, 8] \times 10^4$ bytes are the dynamic ranges for I-frames and P-frames, respectively, while we consider $L \in \{5, 10, 20, 40\}$ as a set up parameter. Moreover, we introduce the possibility to have as input GOP pattern a sequence of GOPs at a certain distance, not necessary consecutive. For example, if $N_G = 4$ and the distance between GOPs is 2 we consider both the sets $m \in \{1, 3\}$ and $m \in \{2, 4\}$ as the indices of the observable GOPs.

Finally, the classifier takes the discrete frame sizes of a certain video as input and estimates the probability distribution of C, i.e., computes the probabilities

$$p_c = \mathrm{P}\left[C = c \,|\, \{I_m, P_{m,n}\}\right] \tag{C.1}$$

that the video belongs to class c, $c \in \{1, 2, 3\}$.

However, due to the poorness of our dataset we actually build three sub-classifiers, one for each class, that have the same structure as the classifier in Fig. C.2. We adopt the "one vs all" technique where every sub-classifier is asked to recognize whether a particular video belongs or not to a given class, i.e., the variable C of Fig. C.2 is substituted with the binary indicator function Y_c, $c \in \{1, 2, 3\}$, which assumes the value 1 if $C = c$ and 0 in all other cases. More precisely, each classifier c estimates the probabilities

$$\begin{cases} p_{c,0} = \mathrm{P}\left[Y_c = 0 \,|\, \{I_m, P_{m,n}\}\right] \\ p_{c,1} = \mathrm{P}\left[Y_c = 1 \,|\, \{I_m, P_{m,n}\}\right] = 1 - p_{c,0}\,. \end{cases} \tag{C.2}$$

The "final" probability distribution of C, is obtained as a combination of the probabilities $p_{c,1}$ derived from the three sub-classifiers, as

$$P\left[C = c \,|\, \{I_m, P_{m,n}\}\right] = \frac{\mathrm{P}\left[Y_c = 1 \,|\, \{I_m, P_{m,n}\}\right]}{\sum\limits_{l=1,2,3} \mathrm{P}\left[Y_l = 1 \,|\, \{I_m, P_{m,n}\}\right]}\,. \tag{C.3}$$

To evaluate the performances of our model, we split the video set into two equally sized partitions. The former is used to *train* the three sub-classifiers deriving the probability distribution of C, while the latter is used as a *test* on videos never processed before. Training and testing sets have approximately the same number of videos. Unfortunately, due to the scarcity and skewness of the available videos (videos of classes 1 and 3 are rare), we adopt two tricks to improve the reliability of the assessed performances.

Firstly, we assume the videos show constant motion level during their entire length, hence all their GOPs have the same sizes distribution. Based on the classifiers input pattern, instead of picking only the first one, we consider all possible GOPs combinations. As an illustrative example: suppose the classifier input pattern is built in such a way it accepts frame sizes from the first and third GOPs, we, then, consider all the possible GOPs combination (n_1, n_2) such that $n_2 = n_1 + 2$, and the stopping criterion is $n_2 \leq N_G$, to avoid biased training and testing set due to very long videos.

Secondly, we repeat the train-&-test partitioning 50 times at random, and finally average the performances obtained over the different experiments.

C.3 Training

We define:

- $\mathcal{X}$ the set of the random variables of the Bayesian network;
- $x_i \in \mathcal{X}$ the i-th random variable of the network represented by a node;
- pa_i the set of parents' nodes of the variable x_i;
- $\theta_{ijk} = p\left(x_i^k | pa_i^j\right)$ the probability $P\left[x_i = k | \, pa_i = j\right]$.

We assume that the prior distribution of θ_{ij} is the Dirichlet distribution, as often assumed in the literature [76], with hyperparameters given by the equivalent sample size ess and the weights τ_{ijk} as

$$\mathrm{P}\left[\theta_{ij}\right] \propto \prod_k \theta_{ijk}^{ess\,\tau_{ijk}-1}, \tag{C.4}$$

where $ess \geq 0$ and $\sum_k \tau_{ijk} = 1$. From [77], the probability distribution θ_{ijk} is expressed by the formula:

$$\theta_{ijk} = \frac{ess\,\tau_{ijk} + n_{ijk}}{ess + \sum\limits_k n_{ijk}}, \tag{C.5}$$

where n_{ijk} is the number of occurrences of the combination $\{x_i = k, pa_i = j\}$ found in the training set.

The learning procedure as in [77] is obtained as a maximization problem of the whole network's variables' entropy over the training set, and can be formalized as

$$\arg\max_{\tau_{ijk}} \; -\sum_{i,j,k} \theta_{ijk} \log \theta_{ijk} \tag{C.6a}$$

$$\text{s.t.} \;\; \theta_{ijk} = \frac{ess\, \tau_{ijk} + n_{ijk}}{ess + \sum_k n_{ijk}} \tag{C.6b}$$

$$\theta_{ijk} \geq 0 \,, \; \sum_k \theta_{ijk} = 1 \tag{C.6c}$$

$$\tau_{ijk} \geq 0 \,, \; \sum_k \tau_{ijk} = 1 \,. \tag{C.6d}$$

As previously mentioned, the network topology of the sub-classifiers is the naive Bayesian network of Fig. C.2 where the class node C is replaced by the random variable Y_c; since the following steps are derived regardless of the particular sub-classifier, we omit for simplicity the dependence on the index c.

We redefine:

- $\mathcal{X} = \{I_m, P_{m,n}, m \in \{1, 2, \ldots, M\}, n \in \{1, 2, \ldots, 15\}, Y_c\}$;
- $\theta_{ijk} = p\left(x_i^k | Y_c^j\right)$ if $x_i = I_m$ or $x_i = P_{m,n}$ (and, consequently, $pa_i = Y_c$, $j = \{0, 1\}$);
- $\theta_{ijk} = \theta_{ik} = p\left(x_i^k\right)$ if $x_i = Y_c$ (and, consequently, $pa_i = \emptyset$).

Based on local optimization criteria, we split the global maximization problem (C.6a) into local optimization problems where every variable x_i and every parent's combination pa_i^j is independent of the others and can be optimized separately. Hence, (C.6a) can be substituted with:

$$\arg\max_{\tau_{ijk}} \; -\sum_{k} \theta_{ijk} \log \theta_{ijk} \,. \tag{C.7}$$

In the learning procedure we use the MATLAB function `fmincon` that performs a series of optimizations to find the parameters τ_{ijk}, and to compute the conditional probability parameters θ_{ijk} through (C.5). Since we work with 3 sub-classifiers in parallel, we actually compute parameters $\theta_{ijk}^{(c)}$, i.e., for every dynamicity class.

C.4 Testing

The inference step is performed using the Maximum A Posteriori (MAP) probability: based on the distribution of the video frames sizes, the network computes the probability distribution over the variable C, and the MAP estimation simply takes the most probable value. Tests are performed over a set of unseen videos and results are averaged over the 50 random combinations as described above. We compute the system performance for each combination of the following set up parameters:

1. the number of quantization intervals L used to discretize the input size frames, $L \in \{5, 10, 20, 40\}$;
2. the number of observable GOPs that varies within $\{1, \ldots, 5\}$;
3. the distance between GOPs that varies within $\{1, 2, 3\}$ (we remark that distance of one corresponds to consecutive GOPs).

The performance metrics we use in this work are precision, recall, and F_1 score [78]. We assess both the performance of the three sub-classifiers separately and the whole classifier as well and we compare our estimator with a Support Vector Machine (SVM) Classifier, built using the Matlab functions available within the Machine Learning Toolbox. We report in Fig. C.3-Fig. C.5 the performance of our Bayesian classifier (left) and of the SVM classifier (right), respectively. From the results there are no outstanding performance improvements in our classifier with respect to SVM, either when varying the number of observable GOPs (Fig. C.3), or the GOPs interdistances (Fig. C.4) or th number of intervals of the quantizers (Fig. C.5), but the common trend in the metrics shows a slight difference in favor of our classifier. Moreover, our classifier shows almost constant results with respect to the number of observable GOPs, in contrast to the SVM which is highly dependent on that configuration.

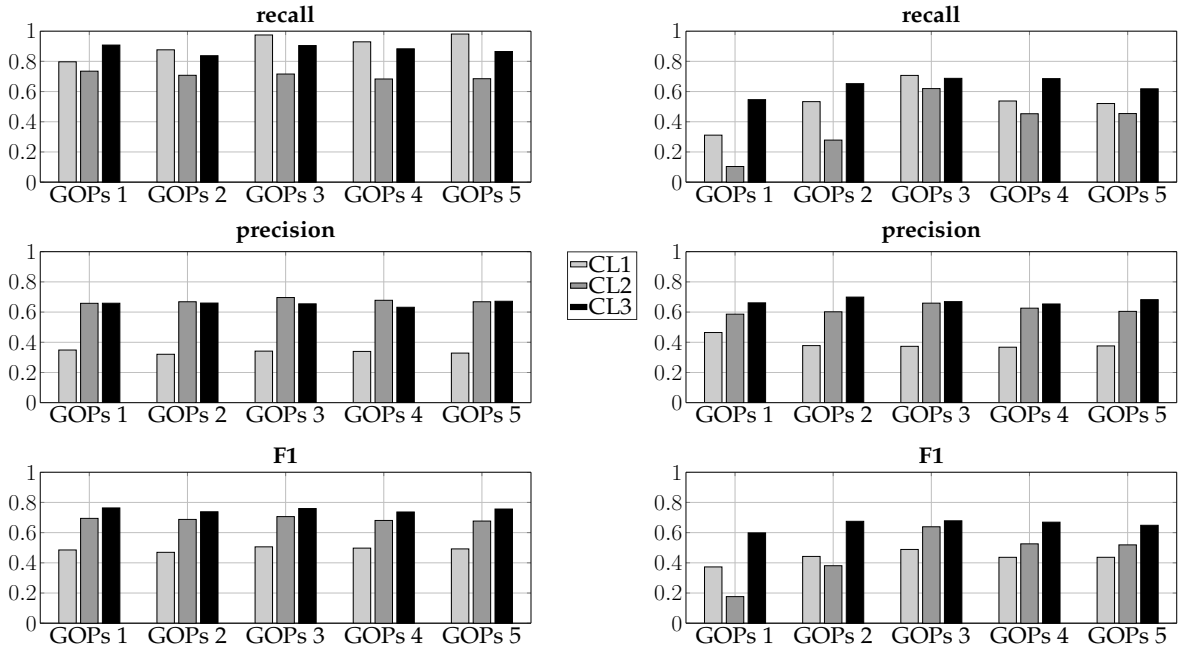

Figure C.3. *Performance of our Bayesian classifier (left) and the SVM classifier (right), with $L = 10$ and distance between GOPs equal to 1; colors refer to the three sub- classifiers, i.e, CL_1 (light gray), CL_2 (dark gray), and CL_3 (black).*

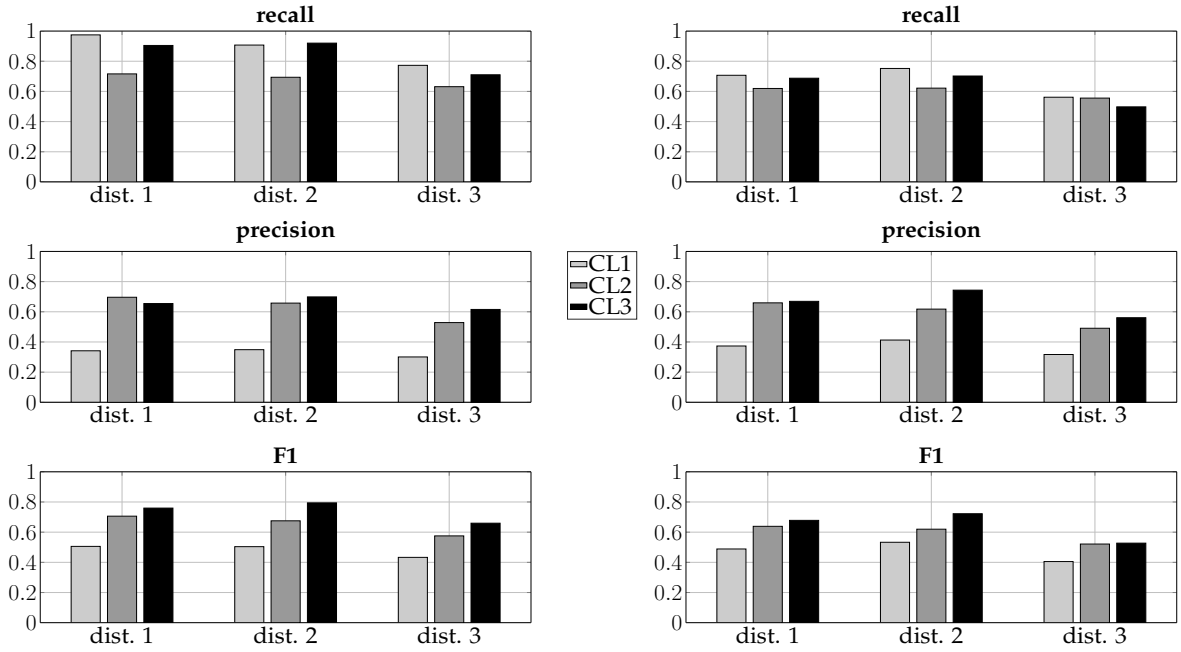

Figure C.4. *Performance of our Bayesian classifier (left) and the SVM classifier (right), with $L = 10$ and observable GOPs equal to 3; colors refer to the three sub- classifiers, i.e, CL_1 (light gray), CL_2 (dark gray), and CL_3 (black).*

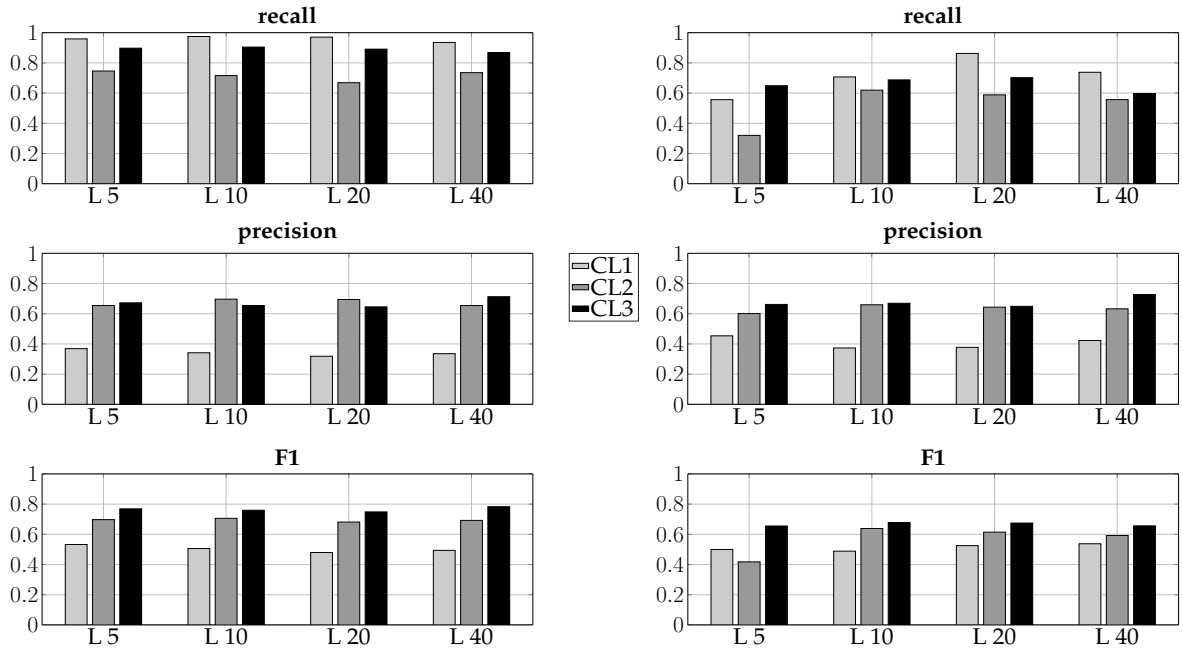

Figure C.5. *Performance of our Bayesian classifier (left) and the SVM classifier (right), with observable GOPs equal to* 3 *and distance between GOPs equal to* 1*; colors refer to the three sub- classifiers, i.e,* CL_1 *(light gray),* CL_2 *(dark gray), and* CL_3 *(black).*

List of Publications

The work presented in this thesis has appeared in the articles reported below.

Journal papers

[J1] F. Guidolin, **I. Pappalardo**, A. Zanella, M. Zorzi, *"Context-Aware Handover Policies in HetNets,"* IEEE Transactions on Wireless Communications, accepted for publication in October 2015.

[J2] **I. Pappalardo**, A. Zanella, M. Zorzi, *"Handover in HetNets: Optimal Performance Analysis,"* submitted to IEEE Wireless Communication Letters.

[J3] **I. Pappalardo**, G. Quer, Bhaskar D. Rao, M. Zorzi, *"Proactive Caching Strategies in HetNets,"* ready for submission to IEEE Transactions on Wireless Communications.

Conference papers

[C1] F. Guidolin, **I. Pappalardo**, A. Zanella, M. Zorzi, *"Context-Aware Handover in HetNets,"* IEEE European Conference on Networks and Communications (EuCNC), June 2014, Bologna, Italy. **BEST STUDENT PAPER AWARD**

[C2] F. Guidolin, **I. Pappalardo**, A. Zanella, M. Zorzi, *"A Markov-based Framework for Handover Optimization in HetNet,"* IEEE Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), June 2014, Piran, Slovenia.

[C3] **I. Pappalardo**, G. Quer, Bhaskar D. Rao, M. Zorzi, *"Caching Strategies in Heterogeneous Networks with D2D, small BS and macro BS communications,"* accepted for publication to IEEE International Conference on Communications (ICC) 2016.

Bibliography

[1] J. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.

[2] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 22–30, June 2011.

[3] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution towards 5g multi-tier cellular wireless networks: An interference management perspective," *CoRR*, vol. abs/1401.5530, 2014. [Online]. Available: http://arxiv.org/abs/1401.5530

[4] T. Korkmaz and M. Krunz, "Multi-constrained optimal path selection," in *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2001, pp. 834–843 vol.2.

[5] S. Podlipnig and L. Böszörmenyi, "A survey of web cache replacement strategies," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 374–398, Dec. 2003. [Online]. Available: http://doi.acm.org/10.1145/954339.954341

[6] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," *White paper*, February 2014.

[7] G. T. 36.839, "Evolved universal terrestrial radio access (e-utra); mobility enhancements in heterogeneous networks (release 11)," Tech. Rep. version 11.0.0, September 2012.

[8] G. T. 36.331, "Protocol specification; radio resource control," Tech. Rep. v.10.4.0, December 2011.

[9] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility management challenges in 3gpp heterogeneous networks," *IEEE Communications Magazine*, vol. 50, no. 12, pp. 70–78, December 2012.

[10] K. Dimou, M. Wang, Y. Yang, M. Kazmi, A. Larmo, J. Pettersson, W. Muller, and Y. Timner, "Handover within 3gpp lte: Design principles and performance," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, Sept 2009, pp. 1–5.

[11] Q. Liao, S. Stanczak, and F. Penna, "A statistical algorithm for multi-objective handover optimization under uncertainties," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, April 2013, pp. 1552–1557.

[12] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in lte-advanced: Key aspects and survey of handover decision algorithms," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 64–91, First 2014.

[13] F. Guidolin, I. Pappalardo, A. Zanella, and M. Zorzi, "A markov-based framework for handover optimization in hetnets," in *13th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET), 2014*, June 2014, pp. 134–139.

[14] ——, "Context-aware handover in hetnets," in *European Conference on Networks and Communications (EuCNC), 2014*, June 2014, pp. 1–5.

[15] ——, "Context-aware handover policies in hetnets," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1–1, 2015.

[16] I. Pappalardo, A. Zanella, and M. Zorzi, "Handover in hetnets: an upper bound analysis," *submitted to IEEE Wireless Communications Letters*, 2016.

[17] Cisco, "Cisco visual networking index: Forecast and methodology, 2014–2019," *White paper*, May 2014.

[18] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, present, and future," *Selected Areas in Communications, IEEE Journal on*, vol. 30, no. 3, pp. 497–508, April 2012.

[19] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, July 2013, pp. 1077–1081.

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

[21] E. Cohen and H. Kaplan, "Exploiting regularities in web traffic patterns for cache replacement," in *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, ser. STOC '99. New York, NY, USA: ACM, 1999, pp. 109–118. [Online]. Available: http://doi.acm.org/10.1145/301250.301281

[22] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *Communications Magazine, IEEE*, vol. 51, no. 4, pp. 142–149, April 2013.

[23] P. Blasco and D. Gunduz, "Learning-based optimization of cache content in a small cell base station," in *Communications (ICC), 2014 IEEE International Conference on*, June 2014, pp. 1897–1903.

[24] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation caching and routing algorithms for massive mobile data delivery," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 3534–3539.

[25] C. Bernardini, T. Silverston, and O. Festor, "Mpc: Popularity-based caching strategy for content centric networks," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 3619–3623.

[26] I. Pappalardo, G. Quer, B. D. Rao, and M. Zorzi, "Caching strategies in heterogeneous networks with d2d, small bs and macro bs communications," *submitted to IEEE International Conference on Communications (ICC) 2016*, 2016.

[27] ——, "Proactive caching strategies in hetnets," *ready for submission to IEEE Transaction on Wireless Communications*, 2016.

[28] M. Peng, D. Liang, Y. Wei, J. Li, and H.-H. Chen, "Self-configuration and self-optimization in lte-advanced heterogeneous networks," *IEEE Communications Magazine*, vol. 51, no. 5, pp. 36–45, May 2013.

[29] D. Lopez-Perez, I. Guvenc, and X. Chu, "Theoretical analysis of handover failure and ping-pong rates for heterogeneous networks," in *IEEE International Conference on Communications (ICC) 2012*, June 2012, pp. 6774–6779.

[30] K. Vasudeva, M. Simsek, and I. Guvenc, "Analysis of handover failures in hetnets with layer-3 filtering," in *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, April 2014, pp. 2647–2652.

[31] C. de Lima, M. Bennis, and M. Latva-aho, "Modeling and analysis of handover failure probability in small cell networks," in *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, April 2014, pp. 736–741.

[32] X. Lin, R. Ganti, P. Fleming, and J. Andrews, "Towards understanding the fundamentals of mobility in cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 4, pp. 1686–1698, April 2013.

[33] K. Kitagawa, T. Komine, T. Yamamoto, and S. Konishi, "A handover optimization algorithm with mobility robustness for lte systems," in *IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications*, Sept 2011, pp. 1647–1651.

[34] Y. Lee, B. Shin, J. Lim, and D. Hong, "Effects of time-to-trigger parameter on handover performance in son-based lte systems," in *16th Asia-Pacific Conference on Communications*, Oct 2010, pp. 492–496.

[35] K. Kitagawa, T. Komine, T. Yamamoto, and S. Konishi, "Performance evaluation of handover in lte-advanced systems with pico cell range expansion," in *IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications*, Sept 2012, pp. 1071–1076.

[36] B. Jeong, S. Shin, I. Jang, N. W. Sung, and H. Yoon, "A smart handover decision algorithm using location prediction for hierarchical macro/femto-cell networks," in *Vehicular Technology Conference (VTC Fall), 2011 IEEE*, Sept 2011, pp. 1–5.

[37] S. Barbera, P. Michaelsen, M. Saily, and K. Pedersen, "Improved mobility performance in lte co-channel hetnets through speed differentiated enhancements," in *IEEE Globecom Workshops*, December 2012, pp. 426–430.

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

[39] N. Zia, S. Mwanje, and A. Mitschele-Thiel, "A policy based conflict resolution mechanism for mlb and mro in lte self-optimizing networks," in *Computers and Communication (ISCC), 2014 IEEE Symposium on*, June 2014, pp. 1–6.

[40] Q. Shen, J. Liu, Z. Huang, X. Gan, Z. Zhang, and D. Chen, "Adaptive double thresholds handover mechanism in small cell lte-a network," in *Wireless Communications and Signal Processing (WCSP), 2014 Sixth International Conference on*, Oct 2014, pp. 1–6.

[41] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Coordinating handover parameter optimization and load balancing in lte self-optimizing networks," in *IEEE 73rd Vehicular Technology Conference*, May 2011, pp. 1–5.

[42] Q. Liao, F. Penna, S. Stanczak, Z. Ren, and P. Fertl, "Context-aware handover optimization for relay-aided vehicular terminals," in *IEEE 14th Workshop on Signal Processing Advances in Wireless Communications*, June 2013, pp. 555–559.

[43] A. Goldsmith, *Wireless Communications*. Cambridge University Press, New York, NJ, 2005.

[44] V. Chandrasekhar and J. Andrews, "Uplink capacity and interference avoidance for two-tier femtocell networks," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 7, pp. 3498–3509, July 2009.

[45] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2002.

[46] J. Lai and N. B. Mandayam, "Minimum duration outages in rayleigh fading channels," *IEEE Transactions on Communications*, vol. 49, no. 10, pp. 1755–1761, Oct 2001.

[47] M. Zorzi, "Outage and error events in bursty channels," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 349–356, Mar 1998.

[48] R. Tanbourgi, S. Singh, J. Andrews, and F. Jondral, "Analysis of non-coherent joint-transmission cooperation in heterogeneous cellular networks," in *IEEE International Conference on Communications*, June 2014, pp. 5160–5165.

[49] K. Fukawa, H. Suzuki, and Y. Tateishi, "Packet-error-rate analysis using markov models of the signal-to-interference ratio for mobile packet systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 6, pp. 2517–2530, July 2012.

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

[51] X. Yan, N. Mani, and Y. Sekercioglu, "A traveling distance prediction based method to minimize unnecessary handovers from cellular networks to wlans," *IEEE Communications Letters*, vol. 12, no. 1, pp. 14–16, January 2008.

[52] M. Garetto, E. Leonardi, and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*, April 2015, pp. 2263–2271.

[53] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, Aug 2014, pp. 649–653.

[54] S. Ioannidis, L. Massoulié, and A. Chaintreau, "Distributed caching over heterogeneous mobile networks," *Queueing Systems: Theory and Applications*, vol. 72, no. 3-4, pp. 279–309, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1007/s11134-012-9297-7

[55] C. Fang, F. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of green information-centric networking: Research issues and challenges," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 3, pp. 1455–1472, third quarter 2015.

[56] A. Araldo, D. Rossi, and F. Martignon, "Cost-aware caching: Caching more (costly items) for less (isps operational expenditures)," *Parallel and Distributed Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[57] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, Nov 2014, pp. 128–133.

[58] T. Wang, L. Song, and Z. Han, "Dynamic femtocaching for mobile users," in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*, March 2015, pp. 861–865.

[59] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, July 2013, pp. 1017–1021.

[60] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium on*, May 2014, pp. 37–42.

[61] K. Poularakis and L. Tassiulas, "Optimal cooperative content placement algorithms in hierarchical cache topologies," in *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, March 2012, pp. 1–6.

[62] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *INFOCOM, 2010 Proceedings IEEE*, March 2010, pp. 1–9.

[63] V. Siris, X. Vasilakos, and G. Polyzos, "Efficient proactive caching for supporting seamless mobility," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on a*, June 2014, pp. 1–6.

[64] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *Information Theory, IEEE Transactions on*, vol. 58, no. 10, pp. 6524–6540, Oct 2012.

[65] M. Taghizadeh, K. Micinski, C. Ofria, E. Torng, and S. Biswas, "Distributed cooperative caching in social wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 12, no. 6, pp. 1037–1053, June 2013.

[66] W. Wu, R. Ma, and J. Lui, "On incentivizing caching for p2p-vod systems," in *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, March 2012, pp. 164–169.

[67] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *Wireless Communications, IEEE Transactions on*, vol. 15, no. 1, pp. 131–145, Jan 2016.

[68] N. Golrezaei, A. Dimakis, and A. Molisch, "Wireless device-to-device communications with distributed caching," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, July 2012, pp. 2781–2785.

[69] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 7, pp. 3665–3676, July 2014.

[70] M. E. J. Newman, "Power laws, pareto distributions and zipfs law," *Contemporary Physics*, 2005.

[71] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, Mar 1999, pp. 126–134 vol.1.

[72] Y. Crama and P. L. Hammer, *Boolean functions : theory, algorithms, and applications*. Cambridge University Press, 2011.

[73] E. Boros and P. L. Hammer, "Pseudo-boolean optimization," *Discrete Appl. Math.*, vol. 123, no. 1-3, pp. 155–225, Nov. 2002. [Online]. Available: http://dx.doi.org/10.1016/S0166-218X(01)00341-9

[74] C. Bielza and P. Larrañaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Surveys*, vol. 47, no. 1, pp. 5:1–5:43, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2576868

[75] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997. [Online]. Available: http://dx.doi.org/10.1023/A%3A1007413511361

[76] P. Walley, "Inferences from multinomial data: Learning about a bag of marbles," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996. [Online]. Available: http://www.jstor.org/stable/2346164

[77] C. de Campos and Q. Ji, "Improving bayesian network parameter learning using constraints," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec 2008, pp. 1–4.

[78] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2005, pp. 345–359.