Linköping Studies in Science and Technology
Dissertations No. 1116

Radio Network Planning and Resource Optimization:

Mathematical Models and Algorithms for UMTS, WLANs, and Ad Hoc Networks

Iana Siomina

Linköpings universitet
INSTITUTE OF TECHNOLOGY

Department of Science and Technology
Linköping University, SE-601 74 Norrköping, Sweden

Norrköping 2007

Linköping Studies in Science and Technology, Dissertations No. 1116

Radio Network Planning and Resource Optimization:
Mathematical Models and Algorithms for UMTS, WLANs, and Ad Hoc Networks

Iana Siomina

Image on front cover was designed by Iana Siomina.

ISBN 978-91-85831-44-9
ISSN 0345-7524
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-9158

Printed by LiU-Tryck, Linköping, Sweden, 2007

Abstract

The tremendous popularity of wireless technologies during the last decade has created a considerable expansion of wireless networks both in size and use. This fact, together with a great variety of mobile devices and numerous different services that are becoming increasingly resource-demanding, have attracted the attention of many researchers into the area of radio resource planning and optimization. Due to network complexity, these tasks require intelligent, automated approaches that are able to deal with many factors in order to enable design of high capacity networks with a high service quality at the lowest possible cost. This is a perfect application of optimization theory.

In this thesis, mathematical optimization is considered as the main approach to designing and improving the performance of wireless networks such as Universal Mobile Telecommunications System (UMTS), Wireless Local Area Networks (WLANs) and ad hoc networks. Due to different underlying access technologies, the optimization goals, design parameters and system limitations vary by network type. Therefore, the goals of the presented work are to identify a relevant optimization problem for each type of network, to model the problem and to apply the optimization approach in order to facilitate wireless network planning and improve radio resource utilization.

The optimization problems addressed in this thesis, in the context of UMTS networks, focus on minimizing the total amount of pilot power which, from the modeling point of view, is not just an amount of power consumed by a certain type of control signal, but also an indicator of the interference level in the network and means of controlling cell coverage. The presented models and algorithms enable flexible coverage planning and optimization of pilot power and radio base station antenna configuration in large networks.

For WLANs, in the first part of the study, the access point placement and the channel assignment problems are considered jointly to maximize net user throughput and minimize co- and adjacent channel interference and contention. The second part of the study addresses the contention issue and involves, among the other decisions, optimization of access point transmit power.

Due to the dynamic and infrastructureless nature of ad hoc networks, static resource planning is less suitable for this type of network. Two algorithmic frameworks which enable dynamic topology control for power-efficient broadcasting in stationary and mobile networks are presented. In both frameworks, the performance of the presented algorithms is studied by simulations.

Keywords: planning, optimization, wireless networks, radio resources, UMTS, WLAN, ad hoc, pilot power, access point location, channel assignment, power allocation, power-efficient broadcast.

Populärvetenskaplig Sammanfattning

Trådlösa nätverk har under det senaste årtiondet blivit enormt populära. I kombination med att det finns mängder av mobila enheter och resurskrävande tjänster har den snabba utbyggnaden av nya nätverk medfört ökad forskningsverksamhet inom planering och optimering av radioresurser. Trådlösa nätverk är väldigt komplexa och för att utforma dessa så att de har hög kapacitet och pålitlighet till en låg kostnad krävs intelligenta tillvägagångssätt som tar hänsyn till flertalet faktorer; detta utgör en perfekt tillämpning för optimeringsteori.

I den här avhandlingen används huvudsakligen optimering för att designa och förbättra trådlösa nätverk som Universal Mobile Telecommunications System (UMTS), Wireless Local Area Networks (WLANs) och ad hoc-nätverk. Dessa nyttjar olika åtkomsttekniker, varför aspekter som optimeringsfokus, designparametrar och systembegränsningar varierar mellan nätverken. Syftet med avhandlingen är att identifiera och presentera relevanta optimeringsproblem för olika trådlösa nätverkstyper, modellera problemen och därefter tillämpa optimering för att underlätta planering samt effektivisera nyttjandet av radioresurser.

Optimeringsproblemen som berör UMTS-nätverk handlar om att minimera den effekt som åtgår för utsändning av en speciell signalsekvens som bland annat används för att skatta egenskaperna hos en kommunikationskanal. De framtagna modellerna och algoritmerna möjliggör optimering av utsänd effektnivå för dessa signaler, flexibel planering av täckningsområden samt reglering av basstationers antennkonfiguration i större nätverk.

Arbetet gällande WLAN handlar dels om hur man lämpligen placerar ut accesspunkter vars uppgift är att sammankoppla närliggande enheter till ett nätverk. Detta innefattar även hur dessa accesspunkter bör tilldelas olika kanaler i syfte att maximera dataöverföring samt minimera olika störningsfaktorer. Dessutom presenteras en studie som bland annat fokuserar på optimering av utsändningseffekt hos utlokaliserade accesspunkter.

Statisk resursplanering är olämpligt för ad hoc-nätverk, som karaktäriseras av föränderlighet och avsaknad av fast infrastruktur. I avhandlingen presenteras två algoritmer för dynamisk topologikontroll som kan nyttjas för att uppnå energieffektiv utsändning i såväl stationära som mobila nätverk. Algoritmerna är utvärderade genom simuleringar.

Acknowledgments

I wish to express my deepest and sincere gratitude to my supervisor Di Yuan, Associate Professor and the head of Mobile Telecommunications group at the Department of Science and Technology (ITN), for his excellent guidance throughout the four years I spent at Linköping University and continuously challenging me to generate new ideas. It has been a great pleasure for me to work with such an excellent researcher and extraordinary person, from whom I tried to learn as much as I could. I also wish to thank Professor Peter Värbrand, my supervisor and the head of our department, for offering me this PhD position and his support and encouragement throughout these four years, as well as for always being open to new ideas and ready to help.

I am very grateful for the financial support I received enabling my research during the four years provided by Center for Industrial Information Technology (CENIIT), Linköping Institute of Technology, under project "Optimal Design och Effektive Planering av Telekommunikationssystem". I also appreciate the financial support I received during the last two years from Swedish Research Council (Vetenskapsrådet) within two projects: "Mathematical Optimization in the Design of UMTS Networks" and "Energy-efficient Broadcasting in Mobile Ad Hoc Networks: Distributed Algorithms and Performance Simulation".

I also wish to thank my current manager at Ericsson Research, Johan Lundsjö, and the SPIRIT project for financially supporting my travel to two recent conferences: the 5th IEEE Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2007) and the 8th IEEE Intl. Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007).

My research work on UMTS networks definitely benefitted from technical discussions with Dr. Fredrik Gunnarsson and his colleagues at Ericsson Research in Linköping, who are also acknowledged for providing a test data set. Also, I am very grateful to Dr. Fredrik Gunnarsson, who was the opponent at my Licentiate seminar in March 2005, for his valuable comments and suggestions on the Licentiate thesis.

My special thanks go to the group of the EU project MOMENTUM IST-2000-28088 for making publicly available the data sets for several European cities, which definitely made the results of my work on UMTS network planning and optimization more valuable from the application point of view.

I am very thankful to COST (European Cooperation in the field of Scientific and Technical Research) Action TIST 293 "Graphs and Algorithms in Communication Networks" for a financial support of my Short-Term Scientific Mission, and the optimization group at Zuse Institute of Berlin, in particular Dr. Andreas Eisenblätter and Hans-Florian Geerdes, for hosting the STSM which resulted in a joint paper that won the Best Paper Award in Helsinki at the WoWMoM 2007 Symposium. Also, I wish to thank Hans-Florian for providing his visualization software I used for generating nice figures in Chapter 7. Jaouhar Jemai from Braunschweig Technical University is also acknowledged for generating radio propagation data for a WLAN.

I am grateful to Dr. Peter Broström, Dr. Sandro Bosio, and Dr. Anders Peterson for

their detailed comments and practical suggestions that were very helpful in improving the presentation quality of this thesis. I would also like to thank my colleagues at ITN for a friendly and inspiring atmosphere in the department.

Finally, I would like to express my thanks to my family for their love and continued care, support and encouragement, and to my friends for their belief in me and being near.

Norrköping, September 2007
Iana Siomina

Contents

List of Tables

List of Figures

List of Algorithms

Abbreviations

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
ACLR	Adjacent Channel Leakage power Ratio
ACP	Adjacent Channel Protection
ACS	Adjacent Channel Selectivity
AGW	Active Gateway
AICH	Acquisition Indicator Channel
AP	Access Point
BSS	Basic Service Set
CCA	Clear Channel Assessment
CCPCH	Common Control Physical Channel
CDMA	Code Division Multiple Access
CDS	Connected Dominating Set
CIR	Carrier-to-Interference Ratio
CPICH	Common Pilot Channel
CS	Carrier Sense
CSMA	Carrier Sense Multiple Access
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
DCF	Distributed Coordination Function
DECT	Digital Enhanced Cordless Telecommunications
DL	Downlink
DPCH	Dedicated Physical Channel
DS	Distribution System
EDGE	Enhanced Data rates for GSM Evolution
ESS	Extended Service Set
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
GSM	Global System for Mobile Communications
HO	Handover
HSDPA	High-Speed Downlink Packet Access
HSUPA	High-Speed Uplink Packet Access
IBSS	Independent Basic Service Set
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
ISDN	Integrated Services Digital Network

ITU	International Telecommunication Union
KPI	Key Performance Indicator
LAN	Local Area Network
LP	Linear Programming
MAC	Medium Access Control
MAN	Metropolitan Area Network
MCDS	Minimum Connected Dominating Set
ME	Mobile Equipment
MIP	Mixed Integer Programming
MT	Mobile Terminal
NP	Nondeterministic Polynomial-time
OFDM	Orthogonal Frequency Division Multiplexing
OVSF	Orthogonal Variable Spreading Factor
P-CPICH	Primary Common Pilot Channel
PAN	Personal Area Network
PCF	Point Coordination Function
PDA	Personal Digital Assistant
PGW	Passive Gateway
PLMN	Public Land Mobile Network
PrGW	Pruning Gateway
PSTN	Public Switched Telephone Network
RAN	Radio Access Network
RBS	Radio Base Station
RF	Radio Frequency
RGW	Regular Gateway
RRM	Radio Resource Management
RSCP	Received Signal Code Power
S-CPICH	Secondary Common Pilot Channel
SCH	Synchronization Channel
SHO	Soft Handover
STA	Station
STDMA	Spatial Time Division Multiple Access
TDMA	Time Division Multiple Access
TDD	Time Division Duplex
TP	Test Point
TPGW	Tentative Passive Gateway
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
USIM	UMTS Subscriber Identity Module
UTRA	Universal Terrestrial Radio Access
UTRAN	UMTS Terrestrial Radio Access Network
WCDMA	Wideband Code Division Multiple Access
WCDS	Weighted Connected Dominating Set
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

Chapter 1

Introduction

1.1 Radio Network Planning and Resource Optimization

With new wireless communication technologies and the increasing size of radio networks, the tasks of network planning and resource optimization are becoming more and more challenging. This is firstly because the radio resource is scarce these days due to the increasing number of subscribers and the many different types of networks operating within the limited frequency spectrum. Secondly, deploying and operating a large network is expensive and therefore requires careful network dimensioning to ensure high resource utilization. As a consequence, manual network design and tuning for improving radio resource allocation are most likely to fail in current and future networks. This necessitates developing automated tools and optimization algorithms that are able to tackle the difficult task. Furthermore, radio network planning and resource optimization can clearly benefit from the well-established optimization theory due to similarities in the objective-oriented way of approaching a problem, selecting the best solution from a number of possible solutions and dealing with many restrictions. In fact, many of the network planning and resource optimization problems can be viewed as specific applications of classical optimization problems.

In this thesis, optimization is considered as the main approach to designing and improving performance of wireless networks such as Universal Mobile Telecommunications System (UMTS), Wireless Local Area Networks (WLANs) and ad hoc networks. The goal is to identify relevant problems for each of the technologies, formalize the problems, and find reasonable solution approaches. First, however, we will discuss what radio network planning and optimization are about, the type of problems they typically address and the typical optimization techniques that can be utilized to solve these problems.

1.1.1 Planning, Optimization or Both?

The tremendous popularity of wireless networks has attracted the attention of many researchers into planning and optimization of wireless networks and radio resources. *Network planning* refers to the process of designing a network structure and determining network elements subject to various design requirements. Network planning is associated with *network dimensioning* and *detailed planning*, i.e., two network life phases both of which are very important since an implemented plan imposes further hard constraints on network performance. In cellular networks, for example, these constraints are very often associated with *hard capacity*. Network performance (capacity) limitation due to resource exhaustion in a particular situation, e.g., high interference and/or heavy load, is often referred to as *soft capacity* (see, for example, [60]). The goal of *resource planning* is to provide a network with the sufficient amount of resources and ensure its effective utilization, whilst achieving a certain minimum amount of hard capacity is usually a task for network planning.

Traditionally, planning is usually viewed as a static task. However, with heterogeneous

radio environments, the concept of dynamic network reconfigurability (see, for example, [32]) featuring the concept of software defined radio [89] has recently gained popularity. Dynamic network planning, by which the network infrastructure and the network mechanisms are to be defined dynamically, has therefore become an attractive research area.

Network optimization amounts to finding a network configuration to achieve the best possible performance. The goal of *resource optimization* is to achieve the best possible resource utilization. The boundary between the two areas, network optimization and resource optimization, is even tighter than that between network planning and resource planning. Moreover, in practice, resource optimization is very often a part of network optimization. The main difference between the two concepts is that optimization tasks related to network infrastructure are typically associated with network optimization rather than resource optimization.

The applicability of optimization techniques for ensuring good network performance is quite intuitive, both for planning a network and/or radio resources, and for optimizing them during operation and maintenance, provided that a reasonable trade-off between the model complexity and reality can be found. Moreover, due to network complexity, its size, and the necessity of dealing with many factors and control parameters, the planning and optimization tasks are often beyond the reach of a manual approach. As a result, the latest trend is *automated wireless network planning and optimization*, initiated by operators of cellular networks but well spread in industrial and research societies (see, for example, [7, 35, 58, 93]). The trend implies using computer systems to generate network design decisions with the minimum amount of human assistance. Such planning systems can clearly benefit from incorporating different optimization modules implementing optimization algorithms. For automated optimization, optimization algorithms are not just a part of the system but are the core of the system. In addition to making the network design process time-efficient, planning and optimization tools can significantly reduce network deployment, operation and maintenance costs. Some interesting challenges arising in radio network planning and optimization are discussed in [44].

1.1.2 Some Classical Optimization Problems in Radio Network Design and Recent Trends

Due to historical reasons and technological and technical aspects of radio communications, the architecture of wireless networks traditionally has had some infrastructure, although infrastructureless networks are becoming more and more popular nowadays. In early networks, the infrastructure was formed by only one radio base station serving the entire intended area. Because of the simple architecture and low flexibility in terms of radio network configuration, radio network design was more focused on system capability and related technical issues, whereas the network planning itself did not receive much attention.

Network planning became a more challenging task with extended network architectures and more radio base stations. Initially, the task involved only two decisions that had to be made in the planning phase: how many radio base stations were needed and where they should be located. This task has been known as the *radio base station location problem*. In the optimization context, this engineering task, which often also involves cost optimization, can be viewed as a facility location problem, one of the classical problems in the field of mathematical programming (more details will be provided in Section 1.2), but with some more interdependencies and requirements imposed by radio access technology. The problem objective is to find a subset of a given set of candidate locations such that the total cost is minimized and the entire area is covered. Note that if feasible solutions are defined in continuous space and the link quality is not uniquely defined by distance between facilities and users, even the single-facility location problem may be computationally difficult and belong to a class of nondeterministic polynomial-time hard, or $\mathcal{NP}$*-hard* [42], problems. The radio base station location problem has been extensively studied for different types of

cellular networks (see, for example, [8, 9, 39, 40, 71, 85, 102, 106]).

Frequency assignment is another classical problem (or, actually, a family of problems) in radio network planning and optimization. The problem was brought into focus with the introduction of the second generation (2G) cellular networks; in particular, Global System for Mobile Communications (GSM) networks, that use a combination of Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and random access. There have been many variations of the frequency assignment problem (see, for example, [1, 16, 34, 54, 72, 92] and the references therein). The basic problem amounts to finding a frequency assignment which is feasible to assignment constraints and interference constraints. Other examples are the minimum interference problem and the minimum span frequency assignment problem. In the first problem, co-channel and adjacent channel interference are minimized. The objective of the second problem is to minimize the difference between the highest and the lowest frequency used in solution. Frequency assignment is typically viewed as a graph coloring problem (see, for example, [36, 54]).

The radio base station location problem and the frequency assignment problem have been the most studied problems in the context of cellular radio network planning and optimization. Among the other, less studied, problems is *topological network design* where the goal is to design network topology of minimum cost that is able to connect the candidate base stations to a fixed (wired) telephone network [33, 86]. The problem involves decisions not only on locations of radio base stations but also on topology of the wired network, i.e., routers, switches, hubs, gateways, etc..

Numerous problems have been considered for optimizing dynamic behavior of radio networks, with and without infrastructure. Among them, the *power control problem*, in which the power assignment is to be decided such that the network capacity is maximized (see, for example, [20, 87]), has been attracting a lot of attention in research. The problem is particularly interesting for interference-limited networks with a small frequency reuse factor, e.g., networks based on Wideband Code Division Multiple Access (WCDMA).

The *rate control* problem arises in wireless networks with adaptive transmission rates. Typically rate control is used to adapt traffic to varying channel conditions, network load, and different QoS requirements while maximizing throughput (see, for example, [76]). When the channel state varies a lot, the objective for delay-sensitive traffic can also be, for example, to minimize the transmission delay and the number of rate switchings in the network [96].

The *scheduling problem* has been studied a lot for wireless ad hoc and sensor networks based on TDMA and Spatial TDMA (STDMA) [15, 41, 51, 98, 100], and now it is getting a new spin with developing networks using Orthogonal Frequency Division Multiplexing (OFDM) and/or adopting adaptive rate mechanisms which allow for more flexible scheduling, e.g., IEEE 802.11 WLANs, WiMAX, and HSDPA [4, 18, 48, 79, 101].

Power control, rate control, and scheduling are also very often considered for joint optimization (see, for example, [75, 91]). Any of these problems can also be combined with frequency or channel assignment (see, for example, [17]).

Due to specific network properties like dynamic topology and no infrastructure, ad hoc and sensor networks give rise to such problems as *virtual topology design* (clustering [6], virtual backbone formation [90], etc.) and *energy efficient routing* problems. Topology design in ad hoc and sensor networks often involves finding a minimum connected dominating set, i.e., solving an optimization problem which is not only $\mathcal{NP}$-hard but also difficult to approximate [82]. Energy efficient routing can be implemented in many different ways. For broadcast, for example, it is very common to utilize a virtual topology such that only a subset of nodes can rebroadcast but all nodes in the network will receive the messages. The objective can be to maximize network lifetime [38] or to construct a minimum-power broadcast tree [31]. Another technique that can make broadcasting and multicasting in ad hoc networks energy efficient is network coding. The minimum energy broadcast/multicast problem can be viewed in this case as a cost minimization problem with linear edge-based pricing, where the edge

prices represent the energy-per-bit of the corresponding physical broadcast links and an edge corresponds to a link in a graph model [111].

Computational complexity of models is an important factor that has often a direct effect on its application. A detailed but intractable mathematical model can be sometimes as useless from the practical point of view as an oversimplified but easily solvable model. This fact has to be considered when developing models. Unfortunately, in most cases models developed for radio network planning and optimization tend to be difficult to solve. In fact, all the optimization problems (at least in their general forms) that have been discussed so far belong to the class of $\mathcal{NP}$-har problems. This means that the existence of polynomial-time algorithms (polynomial with respect to problem size) that can solve these problems is very unlikely, although this has not yet been proven. Moreover, the problem instances tend to be very large in realistic scenarios. This results in that manual parameter adjustment in such networks becomes a tedious task making even more challenging network planning and optimization of radio resources. In such situations exact algorithms typically do not help and even obtaining good approximations can be very difficult. Furthermore, very simple strategies based on one-at-a-time parameter manipulation are not very effective either. Thus, the importance of designing efficient optimization algorithms is not only in solving the problems but also in contributing to the problem application area and thus strengthening the link between theory and practice.

To this point, we have presented a number of typical problems that have been studied in the context of radio network planning and optimization. Although the basic building blocks of models have remained the same (i.e., the classical optimization models are typically adopted by many applications), the optimization problems and modeling approaches have been changing over time in line with technological and scientific trends. Thus, we can distinguish between several major steps in the history of network modeling and optimization. In the first step, the modeling approach mainly followed the trend of oversimplification of reality. Although this was a very important step for establishing relation between the optimization theory and network design, the necessity of more realistic models and practically applicable results gave rise to a new trend — *joint optimization* of several network aspects. This approach has clearly become superior to the previously used sequential optimization approach that exploits model simplicity.

The next trend has been *cross-layer design and optimization* (see, for example, [64, 80]) which actually extends the concept of joint optimization by considering the information flows across the network layers to enable solutions that are globally optimal to the entire system and thus facilitating the optimal layer design. This has been an important step towards decreasing the gap between the optimization state-of-the-art and modeling realism. The most recent trend in network optimization is considering *layering as optimization decomposition* [21, 63] by which the overall communication network is modeled by a master problem where each layer corresponds to a decomposed subproblem, and the interfaces among layers are quantified as functions of the optimization variables coordinating the subproblems. Although the problem decomposition techniques have been widely used for planning and designing networks for quite some time, the new concept is important by itself since it focuses on network architecture but at the same time facilitates distributed control and cross-layer resource allocation. The main difference between the cross-layer design and optimization and the concept of layering as optimization decomposition is that the latter provides also insights into network architecture and layering.

1.2 Mathematical Programming as an Optimization Tool

The term "mathematical programming" refers to a planning process that allocates resources in the best possible, or optimal, way minimizing the costs and maximizing the profits. The mathematical programming approach is to construct a mathematical model, or *mathematical*

program, to represent the problem. In a mathematical model, variables are used to represent decisions, and the quality of decisions is measured by the objective function. Any restrictions on the values of decision variables are expressed by equations and inequalities. A good introduction to the subject of mathematical programming can be found, for example, in [55].

1.2.1 Linear Programming

One of the most important areas of mathematical programming is *linear programming* (LP). The main precursor to LP is considered to be the work published by Leonid Kantorovich in 1939 [65] which laid out the main ideas and algorithms of linear programming. The latter was viewed as a tool for economic planning. The key assumption of LP is that all functions in the model, i.e., objective function and constraint functions, are linear and all variables are continuous. If all or some of the variables are constrained to be integers, the problem is a subject of study for linear *integer programming* and *mixed-integer programming* (MIP), respectively. Mathematical programming problems in which some of the constraints or the objective function are nonlinear are studied by *nonlinear programming*, which is beyond the scope of the thesis.

In a compact way, a typical minimization LP formulation can be represented as follows,

$$\min\{c^{\mathrm{T}}x : Ax \geq b, x \in \mathbb{R}_+\} \ ,$$

where c is a row vector of costs, x is a column vector of nonnegative real variables, A is a matrix, and b is a column vector. (Converting a maximization linear program to the formulation above is straightforward.) The set of feasible solutions to the system of linear inequalities defines a convex polyhedron. One of the commonly used methods for solving linear programs is the simplex method, originally proposed by George Dantzig [27, 28]. The *simplex method* utilizes the concept of a *simplex*[1] and the idea that in the case of closed convex polyhedron the optimum occurs either at a vertex of the polyhedron (and is then unique) or on its edge or face (and is then non-unique). The method finds the optimal solution by moving along the edges of the polyhedron from one vertex to another, adjacent, vertex such that the objective function value does not worsen. In each iteration, a simplex is specified by the set of dependent (basic) variables. A pivot rule is used to decide on the next move if there are several alternatives.

In contrast to the simplex method's approach focusing on extreme feasible solutions on the boundary of the feasible region, Karmarkar's algorithm is an *interior-point algorithm* that cuts through the interior of the feasible region to reach an optimal solution [66]. The algorithm uses transformations from projective geometry to determine the direction for movement. The method is efficient for very large LPs since it is a polynomial-time algorithm. The reason is that, although each iteration of the Karmarkar's algorithm is typically computationally very costly, only a small number of iterations are needed to reach an optimal solution. Simplex and interior-point methods are the important parts of state-of-the-art solvers in any LP software. More information on the interior-point and simplex methods can be found, for example, in [99, 109]. For getting an extensive background on linear programming in general, interested readers are referred to [14, 23].

1.2.2 Integer and Mixed-Integer Programming

Integer programming and mixed-integer programming extend LP to deal with integrality constraints. The focus of integer programming is on integer problems where decision variables may only have integer values. If only some of the variables are required to have integer

[1] A simplex (n-simplex) is a convex hull of a set of $n + 1$ affinely independent vectors in n-dimensional space.

values, the model is referred to as a mixed-integer programming model. Below is a typical minimization MIP formulation,

$$\min\{c^{\mathrm{T}}x + h^{\mathrm{T}}y : Ax + Gy \geq b, x \in \mathbb{R}_+, y \in \mathbb{Z}_+\} ,$$

where x, c, b, and A are as previously defined, y is a column vector of nonnegative integer variables, h is a row vector of costs of y-variables, and G is a matrix. Note that a pure integer program is the special case of the formulation above with no x-variables. An integer program where all variables are binary is called *0-1* or *binary integer program*.

An area closely related to mathematical programming is *combinatorial optimization* which studies problems involving finding the best (with respect to a given objective) solution out of a discrete set of feasible solutions. A combinatorial optimization problem can often be formulated as an integer or binary integer program. Comprehensive treatments of integer programming and combinatorial optimization are given in [24, 94, 103].

Below are some examples of typical integer and mixed-integer programs that arise in radio network planning. Some variations of these programs will also be used in the thesis.

Set Covering. In the classical *set covering problem*, we are given a ground set of M elements and a collection N of subsets of M, the goal is to choose the minimum number of subsets that together cover the entire ground set. The problem has a typical application in radio network coverage planning where the ground set is represented by points that are to be covered, and the collection of subsets represents a set of candidate sites. However, since installation costs as well as operation and maintenance costs typically vary by site, in wireless network planning it is more common to consider the *minimum-cost set covering problem* where a non-negative cost is associated with each subset. The corresponding binary integer programming formulation is given below.

$$\begin{aligned} \min \quad & \sum_{j \in N} c_j x_j & \\ \text{s.t.} \quad & \sum_{j \in N} a_{ij} x_j \geq 1 & i \in M \\ & x_j \in \{0,1\} & j \in N \end{aligned}$$

where x_j is a binary variable that equals one if and only if subset j is selected, and a_{ij} is the element of incidence matrix A such that $a_{ij} = 1$ if and only if element i is covered by subset j. A good survey of algorithms for the problem can be found, for example, in [19].

The minimum-cost set covering problem has been used as a basis for pilot power optimization in Part I. The minimum connected dominating set problem mentioned in the context of ad hoc networks in Part III is also an application of the classical minimum set covering problem.

0-1 Knapsack. Given a set of items N, each with a value and a weight, and the maximum allowed weight W of a knapsack, the decision to be made is selecting a subset of items of maximum total value that satisfies the weight constraint. The problem formulation is as follows.

$$\begin{aligned} \max \quad & \sum_{j \in N} v_j x_j & \\ \text{s.t.} \quad & \sum_{j \in N} w_j x_j \leq W & \\ & x_j \in \{0,1\} & j \in N \end{aligned}$$

where v_j is the value of item j, w_j is the weight of item j, and variable x_j is one if and only if item j is selected. The knapsack problem has been extensively covered in [69, 84]. One

application of the knapsack problem in wireless networks is the problem of maximizing the total network throughput or the number of served users subject to the maximum power budget constraint. In the thesis, the knapsack problem is a subproblem in pilot power optimization with partial coverage.

Facility Location. Let M be a set of possible facility locations and N be a set of clients. Suppose there is a fixed cost f_i of opening a facility at location i, and there is a transportation cost of c_{ij} associated with every facility $i \in M$ and every client $j \in N$. The problem is to decide which facilities to open, and which facility serves each client so as to minimize the sum of the fixed and transportation costs, and every client is assigned to exactly one facility. The problem is known as the *uncapacitated facility location problem*. Note that the problem is similar to the set covering problem except for the addition of the transportation costs. Below is an integer programming formulation of the problem.

$$
\begin{aligned}
\min \quad & \sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij} + \sum_{i \in M} f_i y_i & \\
\text{s. t.} \quad & x_{ij} \leq y_i & i \in M, j \in N \\
& \sum_{i \in M} x_{ij} = 1 & j \in N \\
& x_{ij} \in \{0,1\} & i \in M, j \in N \\
& y_i \in \{0,1\} & i \in M
\end{aligned}
$$

In the formulation, variable y_i is one if and only if facility i is used, and x_{ij} is one if and only if client j is assigned to facility i. Note that in a general formulation of the facility location problem, x-variables are continuous, i.e., $x_{ij} \geq 0, i \in M, j \in N$; however, due to the single assignment property [74] of the uncapacitated facility location problem, a client is always entirely served by the closest facility.

A slight variation of the model presented above is the *capacitated facility location problem* in which each facility i has maximum capacity Q_i and each client $j \in N$ has demand d_j. Thus, in the capacitated facility location problem constraints

$$
\sum_{j \in N} d_j x_{ij} \leq Q_i y_i \qquad i \in M
$$

are typically used. These constraints make redundant the first set of constraints in the formulation of the uncapacitated facility location problem. The redundant constraints are, however, usually kept to make the formulation more efficient (to strengthen its LP relaxation, in particular). Moreover, x-variables are continuous in the classical capacitated facility location problem.

Both types of the facility location problems are used in wireless network planning. One application is the problem of minimizing installation and maintenance costs of base stations in a network providing fixed rate services such that each client gets served. *Maximum k-facility location problem* [25, 26] is a variation of the uncapacitated facility location problem, where the sum of link performance metrics is maximized and the number of facilities is at most k. This type of model has been used for maximizing net user throughput in Chapter 7.

Vertex Coloring. The minimum vertex coloring problem is a problem which aims at finding the minimum number of colors for coloring graph vertices such that adjacent vertices, i.e., vertices connected by an edge in the graph, have distinct colors. The minimum number of the distinct colors is known as the chromatic number of the graph. The problem originates from graph theory, but has been widely considered in integer programming and used in many applications.

Let us consider a graph $G = (V, E)$ where V denotes the set of vertices, and E is the set of edges. The set of candidate colors is denoted by K. An integer programming formulation of the problem, a compact one although not the most efficient, is presented below.

$$
\begin{aligned}
\min \quad & \sum_{k \in K} y_k & & \\
\text{s.t.} \quad & \sum_{k \in K} x_{ik} = 1 & & i \in V \\
& x_{ik} \leq y_k & & i \in V, k \in K \\
& x_{ik} + x_{jk} \leq 1 & & (i,j) \in E, k \in K \\
& x_{ik} \in \{0,1\} & & i \in V, k \in K \\
& y_k \in \{0,1\} & & k \in K
\end{aligned}
$$

where x_{ik} equals one if and only if vertex i is assigned color k, and variable y_k is one if and only if color k is assigned to at least one vertex in the graph. The problem and its variations (e.g., in [1, 36, 34, 72]) have been used to model frequency-assignment-type problems in wireless networks such as channel assignment, code assignment, and time-slot assignment. One example of the problem variations is the minimum-interference frequency assignment problem which allows for assigning the same or adjacent (distinct, but close to each other on a given color scale) colors, but penalizes such assignments for each edge. The objective is then to minimize the sum of all penalties. In this thesis, the problem has been used for modeling channel assignment in WLANs.

All the optimization problems that have been discussed in this section are $\mathcal{NP}$-hard (see, for example, [42, 67, 82]). This fact must be considered when designing solution approaches and implementing the algorithms.

1.2.3 Solution Methods and Techniques for Integer and Mixed-Integer Programs

In general, integer programs are much harder to solve than LPs. The integrality constraints in MIPs are therefore also often considered as the complicating part. The simplest approach to tackle a pure integer programming problem is to explicitly enumerate all possibilities (if they are finite). However, only the smallest instances could be solved by such an approach due to the effect known as *combinatorial explosion* that describes the rapidly accelerating increase in the number of combinations with the increased number of control parameters. In this section, we present the most common approaches used to tackle integer programs and MIPs.

Relaxation and Bounding

Relaxation is an important mathematical programming approach which is used to replace a "difficult" optimization problem by a simpler one by either removing some of the constraints or substituting them with other more easily handled constraints. For a minimization problem, the solution obtained from the relaxation of the original problem gives a lower bound on the optimal solution to the original problem. For maximization, relaxation gives an upper bound. The bound obtained from the relaxation is called a *dual bound*. In minimization problems, any feasible solution is an upper bound on the integer optimum (a lower bound in maximization problems). This bound is also called a *primal bound*. The bounds can be weak or strong depending on how far they are from the optimal solution to the original problem. The relative gap, called *duality gap*, between the primal and the dual bounds is usually used to estimate the quality of the obtained feasible solution. The gap depends on the efficiency of relaxation, solution algorithms, problem size and its complexity.

One of the common relaxation techniques is *LP-relaxation* which involves solving the original problem without integrality constraints, i.e., by treating all variables as continuous variables. A straightforward approach for generating an integer solution from an LP-solution is *rounding* of non-integer values in the resulting LP solution. The drawbacks of this approach are that the solution obtained by rounding is not necessarily feasible and/or it may be far from optimal.

Another relaxation technique is *Lagrangian relaxation* [37], which is a base of one of the algorithms presented in Chapters 3 and 4. Lagrangian relaxation uses the idea of relaxing some constraints by bringing them into the objective function with associated Lagrange multipliers. In a minimization problem, *Lagrangian dual* refers to the problem of maximizing the dual bound with respect to the Lagrange multipliers. Properties of the Lagrangian dual can be found in [62], and its application to integer programming was explored in [43]. *Subgradient method* is very often used to solve the Lagrangian dual and is also applied in Part I of the thesis. By this method, the subgradient direction is obtained by minimizing all the subproblems and then the multipliers are updated along the subgradient direction (see, for example, [5]). Motivated by the fact that the subgradient method may be very computationally expensive for large problems, some variations of the subgradient method have been proposed. The examples are the *interleaved subgradient method* [68], which minimizes only one subproblem per iteration to obtain a direction and then updates the multipliers, and the *surrogate subgradient method* [112], which utilizes the idea that only near optimization of one subproblem is necessary to obtain a proper direction. Other approaches, such as the *analytic center cutting-plane method* [47], *augmented Lagrangian algorithms* (see, e.g., [78]), and *bundle methods* [56, 105], have also been proposed.

Other examples of relaxation techniques are group or modular relaxations (e.g., in [50] and Chapter II.3 in [94]), and surrogate relaxations [45].

Problem Decomposition

Problem decomposition is an approach exploiting the model structure to decompose the problem into smaller and easier-to-solve subproblems. The classical decomposition methods are Lagrangian decomposition, Dantzig-Wolfe decomposition, and Benders decomposition.

By *Lagrangian decomposition* [53], also known as variable splitting, a set of copies of the original variables is introduced for a subset of constraints, and then Lagrangian duality is applied by relaxing the constraints that set equivalence between the original variables and the copies.

In *Benders decomposition* [13], problem variables are partitioned into two sets, master problem variables that are typically complicating variables (e.g., integer variables in a MIP) and subproblem variables. The Benders algorithm iteratively solves a master problem, which assigns tentative values for the master problem variables, and a subproblem obtained by fixing the master problem variables to the tentative values. Solutions to the subproblems are used for generating inequalities (see, for example, [22]) that cut off non-optimal assignments, called Benders cuts, that, being added to the master problems which is to be then resolved, narrow down the search space of the master problem variables.

The key idea of *Dantzig-Wolfe decomposition* [30] is to reformulate the problem by substituting the original variables with a convex combination of the extreme points of a substructure of the formulation. The resulting problem formulations consists of subprograms (slave programs) corresponding to its independent parts and a master program that ties together the subprograms. When solving the master problem, *column generation* can be used to deal with a large number of variables. The main principle of column generation algorithms is to never list explicitly all of the columns (extreme points) of the problem formulation, but rather to generate them only as "needed". Note that for LP, applying Dantzig-Wolfe decomposition is the same as applying Benders decomposition to the dual problem. Originally, Dantzig-Wolfe decomposition was intended for solving LPs, but later it has also been adapted for integer

and mixed-integer programs (e.g., [108]). In Chapter I, a column generation algorithm is embedded into an iterative rounding procedure to obtain primal and dual bounds on the optimal solution to the pilot power optimization problem.

A discussion of modeling with the specific purpose of solving the models with decomposition techniques can be found in [61].

Cutting Planes

Integer programming algorithms often utilize the concept of *cutting planes* (e.g., [29, 49, 94]). Cutting planes remove part of the feasible region of the LP realaxation without removing integer solution points. The basic idea behind a cutting plane is that the optimal integer point is close to the optimal LP solution. Consequently, constraints are added to force the non-integer LP solution to be infeasible without eliminating any integer solutions.

A classical cutting plane approach makes use of *Gomory cuts* [49]. The cuts are used in conjunction with the simplex method and can be generated from any LP solution. Another example is *lift-and-project cut* [10] the idea of which is to consider the integer programming model in a higher dimension (lifting) and to find valid inequalities that, being projected back to the original space, result in a tighter formulation. Very strong cuts can often be generated taking into account the problem structure (e.g., lifted cover inequalities for 0-1 knapsack problem [107]). A good survey of cutting plane approaches for integer and mixed-integer programming is presented in [83].

Branch-and-Bound and Its Extensions

An important technique in integer and mixed-integer programming, *branch-and-bound*, is based on the divide-and-conquer principle that was originally presented in [77] but currently has a lot of extensions (e.g., some more recent experiments with various branch-and-bound strategies can be found in [81]). The branch-and-bound technique is a procedure by which at each iteration a subproblem (the original problem in the first iteration) is subdivided into smaller subproblems by partitioning the set of feasible solutions into smaller subsets. This can be done by restricting the range of the integer variables (for binary variables, there are only two possible restrictions: setting the variable to either 0 or 1). In general, with respect to a variable with lower bound l and upper bound u, the problem will be divided into two subproblems with ranges l to q and $q + 1$ to u, respectively. LP relaxation can be used to obtain bounds on the optimal integer solution. If the optimal solution to a relaxed problem is (coincidentally) integral, it is an optimal solution to the subproblem, and the value can be used to terminate searches of subproblems whose lower bounds are higher. Conquering is done by bounding how good the best solution in the subset can be, and discarding the subset if its bound indicates that it cannot possibly contain an optimal solution for the original problem.

A technique by which cutting planes are embedded into a branch-and-bound framework, is known as *branch-and-cut* [88]. For branch and cut, the lower bound can be, for example, provided by the LP relaxation of the integer program. The optimal solution to this linear program is at an extreme point (vertex) of the feasible region. If the optimal solution to the LP is not integral, this algorithm searches for a constraint which is violated by this solution, but is not violated by any optimal integer solutions, i.e., a cutting plane. When this constraint is added to the LP, the old optimal solution is no longer valid, and so the new optimal solution will be different, potentially providing a better lower bound. Cutting planes are added iteratively until either an integral solution is found or it becomes impossible or too expensive to find another cutting plane. In the latter case, a traditional branch operation is performed and the search for cutting planes continues on the subproblems.

Branch-and-price [11] combines branch-and-bound with column generation. This method is used to solve integer programs where there are too many variables to be handled efficiently

all together. Thus, only a subset of variables is maintained and columns are generated as needed while solving the linear program. Columns with profitable reduced costs are added to the LP relaxation; if no such column exists, the solution is optimal. If the LP solution does not satisfy the integrality constraint, branching is applied.

Dynamic Programming

Another technique often used in integer programming is *dynamic programming* [12] which provides a systematic procedure for determining the optimal sequence of interrelated decisions. The dynamic programming method can be applied to problems that have optimal substructure, i.e., optimal solutions of subproblems can be used to find the optimal solution of the overall problem. For example, the shortest path from one vertex in a graph to another one can be found by first computing the shortest path to the goal from all adjacent vertices, and then using the result to successfully pick the best overall path.

The main idea of the algorithm is that the problem can be broken into smaller subproblems that can be recursively solved to optimality. The found optimal solutions can then be used to construct an optimal solution for the original problem. The subproblems themselves are solved by dividing them into sub-subproblems, and so on, until a simple and easy-to-solve case is reached.

Since the optimal solution is calculated recursively from the optimal solutions to slightly different problems, an appropriate recursive relationship for each individual problem needs to be formulated. On the other hand, if such a recursion exists, we obtain great computational savings over using exhaustive enumeration to find the best combination of decisions, especially for large problems.

Heuristics

Because most of practical problems and many interesting theoretical problems are $\mathcal{NP}$-hard, heuristics and approximation algorithms (the latter is discussed in the next section) play an important role in applied integer programming. Such algorithms are used to find suboptimal solutions when the time or cost required to find an optimal solution to the problem would be very large.

A *heuristic* is typically a simple intuitively designed procedure that exploits the problem structure and does not guarantee an optimal solution. A *meta-heuristic* ("meta" means "beyond") is a general high-level procedure that coordinates simple heuristics and rules to find good approximate (or even optimal) solutions to computationally difficult combinatorial optimization problems. A meta-heuristic does not automatically terminate once a locally optimal solution[2] is found.

Greedy heuristics are simple iterative heuristics specifically designed for a particular problem structure. A greedy heuristic starts with either a partial or infeasible solution and then constructs a feasible solution step by step based on some measure of local effectiveness of the solutions. In each iteration, one or more variables are assigned new values by making greedy choices. The procedure stops when a feasible solution is generated. As an extension of greedy heuristics, a large number of local search approaches have been developed to improve given feasible solutions.

Lagrangian heuristics exploit the solution process of the Lagrangian dual in order to obtain feasible solutions to the original problem. Most Lagrangian heuristics proposed so far attempt to make the optimal solution to the Lagrangian relaxation feasible, e.g., by means of a simple heuristic.

Local search [3] is a family of methods that iteratively search through the set of solutions. Starting from an initial feasible solution, a local search procedure moves from one feasible

[2] *Local optimum* is a solution optimal within a neighboring set of solutions; this is in contrast to a *global optimum*, which is the optimal solution in the whole solution space.

solution to a neighboring solution with a better objective function until a local optimum is found or some stopping criteria are met. The next two algorithms, simulated annealing and tabu search, enhance local search mechanisms with techniques for escaping local optima.

Simulated annealing [2, 70, 110] is a probabilistic meta-heuristic derived from statistical mechanics. This iterative algorithm simulates the physical process of annealing, in which a substance is cooled gradually to reach a minimum-energy state. The algorithm generates a sequence of solutions and the best among them becomes the output. The method operates using the neighborhood principle, i.e., a new solution is generated by modifying a part of the current one and evaluated by the objective function (corresponding to a lower energy level in physical annealing). The new solution is accepted if it has a better objective function value. The algorithm also allows occasional non-improving moves with some probability that decreases over time, and depends on an algorithm parameter and the amount of worsening. A non-improving move means to go from one solution to another with a worse objective function value. This type of move helps to avoid getting stuck in local optimum. It has been proved that with a sufficiently large number of iterations and a sufficiently small final temperature, the simulated algorithm converges to a global optimum with a probability close to one [2, 95]. However, with these requirements, the convergence rate of the algorithm is very low. Therefore, in practice it is more common to accelerate the algorithm performance to obtain fast solution approximations.

Tabu search [46] is a meta-heuristic technique that operates using the following neighborhood principle. To produce a neighborhood of candidate solutions in each iteration, a solution is perturbed a number of times by rules describing a move. The best solution in the neighborhood replaces the current solution. To prevent cycling and to provide a mechanism for escaping locally optimal solutions, some moves at one iteration may be classified as tabu if the solutions or their parts, or attributes, are in the tabu list (the short-term memory of the algorithm), or the total number of iterations with certain attributes exceeds a given maximum (long-term memory). There are also aspiration criteria which override the tabu moves if particular circumstances apply.

Genetic algorithms [59, 73, 97] are probabilistic meta-heuristics that mimic some of the processes of evolution and natural selection by maintaining a population of candidate solutions, called individuals, which are represented by strings of binary genes. A genetic algorithm starts with an initial population of possible solutions and then repeatedly applies operations such as crossover, mutation, and selection to the set of candidate solutions. A crossover operator generates one or more solutions by combining two or more candidate solutions, and a mutation operator generates a solution by slightly perturbing a candidate solution. Thus, the population of solutions evolves via processes which emulate biological processes. Introduced by Holland [59], the basic concept is that the strong species tend to adapt and survive while the weak ones tend to die out.

Approximation Algorithms

Approximation is another approach to deal with difficult optimization problems. *Approximation algorithms* are the algorithms that guarantee the quality of the solution and run in polynomial time in the worst (or average for randomized algorithms) case. The performance guarantee can be an absolute value, i.e., the maximum difference between the optimal value and the approximation, but it is more common to specify a relative performance guarantee. The latter implies that the algorithm produces a solution whose value is always within a factor of α of the value of an optimal solution. Note that $\alpha > 1$ for minimization problems and $\alpha < 1$ for maximization problems ($\alpha = 1$ means the existence of an exact polynomial-time algorithm). Such an algorithm is called an α-approximation algorithm. If for every $\epsilon > 0$ there exists an $(1 + \epsilon)$ approximation algorithm (for a minimization problem; $(1 - \epsilon)$ is considered for a maximization problem), the problem admits a polynomial-time approximation scheme (PTAS). The best approximability property of a problem is the existence of

a Fully PTAS (FPTAS), i.e., when for every $\epsilon > 0$ there exists an approximation algorithm which is also polynomial with respect to $1/\epsilon$. The binary knapsack and the maximum facility location problems are the examples. Problems that can be approximated within some, but not every, constant factor α are considered to have a worse approximability property. An example of such problems is the classical minimum facility location problem which is known to be approximable within 2.408 but not within 1.463 and thus does not admit PTAS [52]. A survey of approximation algorithms for $\mathcal{NP}$-hard problems and results for covering and network design problems, among others, can be found in [57].

1.3 Scope and Structure of the Thesis

1.3.1 Thesis Outline and Organization

The remainder of the thesis is organized in three parts that address issues of radio network planning and resource management in three types of wireless networks, namely, UMTS, WLANs, and ad hoc networks. The content of each of the three parts is outlined below.

- **Network Planning and Resource Optimization for UMTS (Part I).** Pilot power management in 3G networks based on Wideband Code Division Multiple Access (WCDMA) is in focus in this part of the thesis. An example of such networks is UMTS that has gained popularity recent years. The transmit power of pilot signals in UMTS does not only affects the power consumption in a cell, but also cell capacity and coverage. Two strategies of assigning pilot power are under investigation, uniform and non-uniform. The objective is to minimize the total pilot power in the network subject to a coverage requirement. Simple heuristic solutions are presented and compared to more advanced modeling and optimization approaches. The basic pilot power optimization problem subject to a full coverage constraint is formulated in Chapter 3 and is then extended with a partial coverage requirement in Chapter 4. Chapter 5 considers a problem of optimizing radio base station configuration where, in addition to pilot power, antenna azimuth and antenna mechanical and electrical tilts are the decision variables. Extensive numerical studies conducted on data sets representing real planning scenarios are discussed.

- **Coverage Planning and Radio Resource Optimization for Wireless LANs (Part II).** The scope of this part is radio resource management (RRM) for Wireless LANs. To improve network performance and to optimize radio resource utilization, a number of optimization models are proposed to enable efficient network planning. Among the planning decisions that can benefit from the proposed models are access point (AP) placement, channel assignment, and AP transmit power adjustment. In Chapter 7, the objectives are to maximize user throughput and to minimize AP coverage overlap. Numerical experiments have been conducted for a real network to compare sequential and integrated optimization approaches. In Chapter 8, the presented model takes also into account contention among user terminals. The chapter also includes a study on lower and upper bounding approaches.

- **Managing Dynamic Power-efficient Broadcast Topologies in Ad Hoc Networks (Part III).** The last part presents two algorithmic frameworks addressing power efficiency of broadcasting in ad hoc networks. Both frameworks aim at designing distributed algorithms for dynamic management of power-efficient broadcast topologies. The first framework focuses on stationary networks (Chapter 10), and the second one is designed for mobile networks (Chapter 11). Two different power-efficiency optimization goals are considered: prolonging network lifetime (for stationary networks) and finding a minimum-power broadcast topology (for mobile networks). Two topology control approaches are studied for static networks: maintaining a dynamic virtual backbone and

dynamic adjustment of transmit power. For mobile networks, only the first topology control approach is considered, but two different pruning schemes are investigated. The four algorithms are presented, analyzed, and studied by simulations.

Each part of the thesis has its own bibliography list and uses the terminology that is specific for the addressed wireless technology. Terminology is typically introduced in the introductory chapter of the corresponding part, but abbreviations may be reused later in other parts of the thesis. A complete abbreviation list is included in the beginning of the thesis.

Mathematical notation is defined separately for each model, unless explicitly stated otherwise. Some notation can be reused within a chapter when the models are connected to each other, but the reader is then given a reference to the place where the notation is originally introduced. Equations, theorems, and propositions are numbered separately in every part.

1.3.2 Contributions

The contributions are summarized below and grouped according to the thesis structure.

- **UMTS networks (Part I)**
  - The pilot power optimization problem for WCDMA networks has been identified and formulated by means of integer programming. Several models are presented in the thesis. The basic formulation addresses a trade-off between the service area coverage and pilot power consumption. One of the model extensions is designed to take into account traffic distribution. The application of the second extension is optimization of radio base station configurations.
  - Efficient solution and lower bounding approaches for the developed models have been designed and implemented to enable effective planning of large networks.
  - Simple ad hoc solutions have been derived for comparison with the solutions obtained by optimization.
  - Numerical studies have been conducted on realistic data sets for several European cities provided by the EU-project MOMENTUM [104]. A suite of Java-Matlab tools has been developed for the data set processing and visualization.
- **Wireless LANs (Part II)**
  - The network planning problems have been identified and modeled taking into account the key aspects of the underlying technology, e.g., user throughput, contention, and co- and adjacent channel interference.
  - Several optimization models have been designed and mathematically formulated to enable optimized AP placement and channel assignment such that the decisions could be taken sequentially and jointly.
  - Numerical studies have been conducted for various scenarios in a multi-floor office environment originated from a real WLAN in an office building of Zuse Institute Berlin (ZIB). The design patterns drawn from the optimal solutions can be used to derive guidelines or effective strategies for WLAN planning and optimization for similar environments.
  - Taking into account uplink communication, an optimization model has been developed for minimizing contention by deciding on AP transmit power levels and channel assignment.
  - A comparative numerical study has been conducted for several lower bounding approaches based on strengthening the channel assignment part of the models.

- **Ad hoc networks (Part III)**
  - Two distributed algorithms have been designed for dynamic topology control in stationary networks with non-adjustable and adjustable transmit power.
  - Two distributed asynchronous algorithms have been designed for dynamic virtual backbone update in mobile networks.
  - A Java-based tool has been developed for simulation and visualization of mobile ad hoc networks. The tool has been used for studying the performance of the proposed power-efficient broadcast algorithms.

1.3.3 Publications

The thesis provides a detailed and structured description of the research work that has been presented in the publications listed in this section. For the readers' convenience, the publications are organized by technology.

- **UMTS networks (Part I)**

  The models for pilot power optimization and some extensions have been presented in detail in

  I. Siomina. Pilot power management in radio network planning for WCDMA networks. Licentiate thesis, Linköping Institute of Technology, Feb. 2005.

  I. Siomina, P. Värbrand, D. Yuan. Pilot power optimization and coverage control in WCDMA mobile networks. To appear in *OMEGA - The International Journal of Management Science*, Special Issue on Telecommunications Applications, 35(6):683–696. Elsevier, Dec. 2007.

  I. Siomina and D. Yuan. Minimum pilot power for service coverage in WCDMA networks. To appear in *ACM/Kluwer Journal of Wireless Networks (WINET)*, 2007.

  I. Siomina, P. Värbrand, and D. Yuan. Automated optimization of service coverage and base station antenna configuration in UMTS networks. In A. Capone and J. Zhang (eds), *IEEE Wireless Communications Magazine, Special Issue on 3G/4G/WLAN/WMAN Planning and Optimisation*, pages 16–25, Dec. 2006.

  The following conference papers cover some of the material presented in Part I.

  I. Siomina and D. Yuan. Pilot power optimization in WCDMA networks. In *Proc. of the 2nd Workshop on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt '04)*, Cambridge, UK, pages 191–199, March 2004.

  I. Siomina and D. Yuan. Pilot power management in WCDMA networks: Coverage control with respect to traffic distribution. In *Proc. of the 7th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '04)*, Venice, Italy, pages 276–282, Oct. 2004.

  I. Siomina and D. Yuan. Optimization of pilot power for service coverage and smooth handover in WCDMA networks. In E. M. Belding-Royer, K. A. Agha, G. Pujolle (eds), *Mobile and Wireless Communications Networks*, Springer, pages 191–202, Oct. 2004.

  I. Siomina. P-CPICH power and antenna tilt optimization in UMTS networks. In *Proc. of IEEE Advanced Industrial Conference on Telecommunications (AICT '05)*, Lisbon, Portugal, pages 268–273, July 2005.

  I. Siomina, P. Värbrand, and D. Yuan. An effective optimization algorithm for configuring radio base station antennas in UMTS networks. In *Proc. of the 64th IEEE Vehicular Technology Conference 2006 Fall (VTC2006-fall)*, Montréal, Sep. 2006.

- **Wireless LANs (Part II)**

  The presented research work on Wireless LANs originated from a research visit within COST Action TIST 293 "Graphs and Algorithms in Communication Networks" (GRAAL) and has been originally presented in the STSM technical report:

  > I. Siomina. Wireless LANs planning and optimization. STSM Technical Report, COST Action TIST 293, Dec. 2005.

  An extension of this work contains some of the material presented in Section 7 and has been published as the following joint paper which won the IEEE WoWMoM 2007 Best Paper Award,

  > A. Eisenblätter, H.-F. Geerdes, and I. Siomina. Integrated access point placement and channel assignment for Wireless LANs in an indoor office environment. In *Proc. of the 8th IEEE Intl. Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '07)*, Helsinki, Finland, June 2007.

  Chapter 8 is based on the material presented in the following conference publication, but provides more detailed theoretical and numerical studies.

  > I. Siomina and D. Yuan. Optimization of channel assignment and access point transmit power for minimizing contention in Wireless LANs. In *Proc. of the 5th IEEE Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '07)*, Limassol, Cyprus, Apr. 2007.

- **Ad hoc networks (Part III)**

  The first algorithmic framework presented in this part (Chapter 10) covers two conference papers on dynamic management a virtual backbone and a paper on dynamic adjustment transmit power in static networks.

  > I. Siomina and D. Yuan. Extending broadcast lifetime in ad hoc networks by distributed and smooth backbone update. In *Proc. of the 3rd IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS '06)*, Vancouver, pages 497–500, Oct. 2006.

  > I. Siomina and D. Yuan. A distributed hybrid algorithm for broadcasting through a virtual backbone in wireless ad hoc networks. In *Proc. of the 6th Scandinavian Workshop on Wireless Ad-Hoc Networks (Adhoc '06)*, Johannesberg, Sweden, May 2006.

  > I. Siomina and D. Yuan. Maximizing lifetime of broadcasting in ad hoc networks by distributed transmission power adjustment. In *Proc. of the 8th International Conference on Transparent Optical Networks (ICTON '06)*, Nottingham, UK, pages 248–252, June 2006.

  The following two conference papers cover the material presented in Chapter 11.

  > I. Siomina and D. Yuan. Managing a dynamic broadcast infrastructure in mobile ad hoc networks through distributed and asynchronous update of a virtual backbone. In *Proc. of the 25th IEEE Military Communications Conference (MILCOM '06)*, Washington, DC, Oct. 2006.

  > I. Siomina and D. Yuan. Managing a broadcast infrastructure in ad hoc networks in presence of mobility: A new algorithmic framework. In *Proc. of the 65th IEEE semiannual Vehicular Technology Conference (VTC2007-Spring)*, Dublin, Ireland, pages 71–75, Apr. 2007.

Bibliography

[1] K. Aardal, S. Van Hoesel, A. M. C. A. Koster, C. Mannino, and A.Sassano. Models and solution techniques for frequency assignment problems. *4OR: A Quarterly Journal of Operations Research*, 1(4):261–317, Dec. 2003.

[2] E. H. L. Aarts and J. H. M. Korst. *Simulated Annealing and Boltzmann Machines*. Wiley, 1989.

[3] E. H. L. Aarts and J. K. Lenstra, editors. *Local Search in Combinatorial Optimization*. Wiley, 1997.

[4] R. Agrawal, R. Berry, J. Huang, and V. Subramanian. Optimal scheduling for OFDMA systems. In *Proc. of 40th Annual Asilomar Conference on Signals, Systems, and Computers*, Oct. 2006.

[5] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[6] J. N. Al-Karaki, A. E. Kamal, and R. Ul-Mustafa. On the optimal clustering in mobile ad hoc networks. In *Proc. of the First IEEE Consumer Communications and Networking Conference (CCNC 2004)*, pages 71–76, Jan. 2004.

[7] S. M. Allen and R. K. Taplin. Automatic design of fixed wireless access networks. *Intl. Journal of Mobile Network Design and Innovation*, 1(1):1–7, Oct. 2005.

[8] E. Amaldi, A. Capone, and F. Malucelli. Discrete models and algorithms for the capacitated location problems arising in UMTS network planning. In *Proc. of the 5th Intl. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIAL-M 2001)*, pages 1–8. ACM Press, July 2001.

[9] E. Amaldi, A. Capone, and F. Malucelli. Optimizing base station siting in UMTS networks. In *Proc. of the 53rd IEEE Vehicular Technology Conference (VTC2001-Spring)*, May 2001.

[10] E. Balas, S. Ceria, and G. Cornuéjols. A lift-and-project cutting plane algoritm for mixed 0–1 programs. *Mathematical Programming*, 58(3):295–324, Feb. 1993.

[11] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. P. Savelsbergh, and P. H. Vance. Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46(3):316–329, May 1998.

[12] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[13] J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–25, 1962.

[14] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[15] P. Björklund, P. Värbrand, and D. Yuan. A column generation method for spatial TDMA scheduling in ad hoc networks. *Ad Hoc Networks*, 2(4):405–418, Oct. 2004.

[16] P. Björklund, P. Värbrand, and D. Yuan. Optimal frequency planning in mobile networks with frequency hopping. *Computers and Operations Research*, 32:169–186, Jan. 2005.

[17] M. Bohge, J. Gross, and A. Wolisz. The potential of dynamic power and sub-carrier assignments in multi-user OFDM-FDMA cells. In *Proc. of the 48th Annual IEEE Global Telecommunications Conference (GLOBECOM 2005)*, Nov. 2005.

[18] M. Bottigliengo, C. Casetti, C.-F. Chiasserini, and M. Meo. Smart traffic scheduling in 802.11 WLANs with access point. In *Proc. of the 58th Vehicular Technology Conference (VTC2003-Fall)*, pages 2227–2231, Oct. 2003.

[19] A. Caprara, M. Fischetti, and P. Toth. Algorithms for the set covering problem. *Annals of Operations Research*, 98(1–4):353–371, Dec. 2000.

[20] D. Catrein, L. Imhof, and R. Mathar. Power control, capacity, and duality of up and downlink in cellular CDMA systems. *IEEE Transactions on Communications*, 52(10):1777–1785, 2004.

[21] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle. Layering as optimization decomposition: Questions and answers. In *Proc. of Military Communications Conference (MILCOM 2006)*, Oct. 2006.

[22] Y Chu and Q. Xia. *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, volume 3011 of *Lecture Notes in Computer Science*, chapter Generating Benders Cuts for a General Class of Integer Programming Problems, pages 127–141. Springer, May 2004.

[23] V. Chvátal. *Linear Programming*. Freeman, New York, 1983.

[24] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver. *Combinatorial Optimization*. Wiley, New York, 1997.

[25] G. Cornuejols, M. Fisher, and G. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23:789–810, 1977.

[26] G. Cornuejols, R. Sridharan, and J. M. Thizy. A comparison of heuristics and relaxations for the capacitated plant location problem. *European Journal of Operational Research*, 50(3):280–297, Feb. 1991.

[27] G. B. Dantzig. Programming in a linear structure. *Econometrica*, 17:73–74, 1949.

[28] G. B. Dantzig. *Linear programming and extensions*. Princeton University Press, Princeton NJ, 1963.

[29] G. B. Dantzig, D. R. Fullkerson, and S. M. Johnson. Solution of a large scale traveling salesman problem. *Operations Research*, 2:393–410, 1954.

[30] G. B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, Jan. 1960.

[31] A. K. Das, R. J. Marks, M. El-Sharkawi, P. Arabshahi, and A. Gray. Minimum power broadcast trees for wireless networks: Integer programming formulations. In *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, pages 1001–1010, 2003.

[32] P. Demestichas, G. Dimitrakopoulos, J. Luo, R. Agusti, E. Mohyledinand, O. Sallent, D. Grandblaise, R. Pintenet, P. Leaves, and K. Moessner. Radio resource management and network planning in a reconfigurability context. In *Proc. of IST Mobile and Wireless Summit 2004*, June 2004.

[33] A. Dutta and P. Kubat. Design of partially survivable networks for cellular telecommunication systems. *European Journal of Operational Research*, 118(1):52–64, Oct. 1999.

[34] A. Eisenblätter. *Frequency assignment in GSM networks: Models, heuristics, and lower bounds*. PhD thesis, Technische Universität Berlin, Berlin, Germany, 2001.

[35] A. Eisenblätter, A. Fügenschuh, E. R. Fledderus, H.-F. Geerdes, B. Heideck, D. Junglas, T. Koch, T. Kürner, and A. Martin. Mathematical methods for automatic optimization of UMTS radio networks. Technical Report D4.3, IST-2000-28088 MOMENTUM, 2003.

[36] A. Eisenblätter, M. Grötschel, and A. M. C. A. Koster. Frequency planning and ramifications of coloring. *Discussiones Mathematicae Graph Theory*, 22(1):51–88, 2002.

[37] M. L. Fisher. The Lagrangian relaxation method for solving integer programming problems. *Management Sience*, 27(1):1–18, Jan. 1981.

[38] P. Floréen, P. Kaski, J. Kohonen, and P. Orponen. Multicast time maximization in energy constrained wireless networks. In *Proc. of Intl. Conference on Mobile Computing and Networking (MobiCom 2003)*, pages 50–58, Sep. 2003.

[39] L. C. P. Floriani and G. R. Mateus. Optimization models for effective cell planning design. In *Proc. of the First Intl. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, 1997.

[40] M. Galota, C. Glaßer, S. Reith, and H. Vollmer. A polynomial-time approximation scheme for base station positioning in UMTS networks. In *Proc. of the 5th Intl. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, pages 52–59. ACM Press, July 2001.

[41] S. Gandham, M. Dawande, and R. Prakash. Link scheduling in sensor networks: Distributed edge coloring revisited. In *Proc. of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2005)*, pages 2492–2501, March 2005.

[42] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[43] A. M. Geoffrion. Lagrangean relaxation for integer programming. *Mathematical Programming Study*, 2:82–114, 1974.

[44] T. Gill. Radio planning and optimisation – the challenge ahead. In *Proc. of the 4th Intl. Conference on 3G Mobile Communication Technologies (3G 2003)*, pages 28–30, June 2003.

[45] F. Glover. Surrogate constraints. *Operations Research*, 16:741–749, 1968.

[46] F. Glover, E. Taillard, and D. de Werra. A user's guide to tabu search. *Annals of Operations Research*, 41:3–28, 1993.

[47] J.-L. Goffin, A. Haurie, and J.-P. Vial. Decomposition and nondifferentiable optimization with the projective algorithm. *Management Science*, 38:284–302, 1992.

[48] J. S. Gomes, M. Yun, H.-A. Choi, J.-H. Kim, J. K. Sohn, and H. I. Choi. Scheduling algorithms for policy driven QoS support in HSDPA networks. In *Proc. of the 26th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2007)*, May 2007.

[49] R. E. Gomory. An algorithm for integer solutions to linear programs. In *Recent Advances in Mathematical Programming*, pages 269–302. McGraw-Hill, New York, 1963.

[50] R. E. Gomory. On the relaxation between integer and non-integer solutions to linear programs. *Proc. of the National Academy of Sciences*, 53:260–265, 1965.

[51] J. Grönkvist. Novel assignment strategies for spatial reuse TDMA in wireless ad hoc networks. *Wireless Networks*, 12(2):255–265, June 2006.

[52] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. In *Proc. of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 649–657, 1998.

[53] M. Guignard and S. Kim. Lagrangian decomposition: A model yielding stronger Lagrangian bounds. *Mathematical Programming*, 39(2):215–228, 1987.

[54] W. K. Hale. Frequency assignment: Theory and applications. In *Proc. of the IEEE*, volume 68, pages 1497–1514, Dec. 1980.

[55] F. S. Hillier and G. J. Lieberman. *Introduction to Mathematical Programming*. McGraw-Hill, 1991.

[56] J.-B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Springer, 2nd part edition edition, Dec. 1993.

[57] D. S. Hochbaum. *Approximation Algorithms for NP-hard Problems*. PWS Publishing, Boston, 1995.

[58] A. Höglund and K. Valkealahti. Automated optimization of key WCDMA parameters. *Wireless Communications and Mobile Computing*, 5(3):257–271, May 2005.

[59] J. H. Holland. *Adaptation in natural and artificial systems*. The University of Michigan Press, Ann Arbor, 1975.

[60] H. Holma and A. Toskala. *WCDMA for UMTS – Radio Access for Third Generation Mobile Communications*. Wiley & Sons, Aug. 2004.

[61] K. Holmberg. Creative modeling: Variable and constraint duplication in primal-dual decomposition methods. *Annals of Operations Research*, 82:355–390, 1998.

[62] H. Everett III. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11:399–417, 1963.

[63] B. Johansson, P. Soldati, and M. Johansson. Mathematical decomposition techniques for distributed cross-layer optimization of data networks. *IEEE Journal on Selected Areas in Communications*, 24(8):1535–1547, Aug. 2006.

[64] M. Johansson and L. Xiao. Cross-layer optimization of wireless networks using nonlinear column generation. *IEEE Transactions on Wireless Communications*, 5(2):435–445, Feb. 2006.

[65] L. Kantorovich. *Mathematical Methods of Organising and Planning Production*. Leningrad University Press, 1939.

[66] N. Karmarkar. A new polynomial time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

[67] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J.W. Tatcher, editors, *Complexity of Computer Computations*, pages 85–103. New York: Plenum, 1972.

[68] C. A. Kaskavelis and M. C. Caramanis. Efficient Lagrangian relaxation algorithms for industry size job-shop scheduling problems. *IIE Transactions*, 30:1085–1097, 1998.

[69] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer, 2005.

[70] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[71] I. Kocsis, I. L. Farkas, and L. Nagy. 3G base station positioning using simulated annealing. In *Proc. of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002)*, pages 330–334, Sep. 2002.

[72] A. M. C. A. Koster. *Frequency Assignment – Models and Algorithms*. PhD thesis, Maastricht University, 1999.

[73] J. R. Koza. Survey of genetic algorithms and genetic programming. In *Proc. of WESCON '95 - Conference Record: Microelectronics, Communications Technology, Producing Quality Products, Mobile and Portable Power, Emerging Technologies*, 1995.

[74] J. Krarup and P. M. Pruzan. The simple plant location problem: Survey and synthesis. *European Journal of Operational Research*, 12:36–81, 1983.

[75] G. Kulkarni, V. Raghunathan, and M. Srivastava. Joint end-to-end scheduling, power control and rate control in multi-hop wireless networks. In *Proc. of the 47th annual IEEE Global Telecommunications Conference (GLOBECOM 2004)*, number 29, Dec. 2004.

[76] Y.-L. Kuo, K.-W. Lai, F. Y.-S. Lin, Y.-F. Wen, E. H.-K. Wu, and G.-H. Chen. Multi-rate throughput optimization for wireless local area network anomaly problem. In *Proc. of the 2nd Intl. Conference on Broadband Networks*, pages 591–601, Oct. 2005.

[77] A. H. Land and A. G. Doig. An automatic method for solving discrete programming problems. *Econometrica*, 28:497–520, 1960.

[78] T. Larsson and D. Yuan. An augmented lagrangian algorithm for large scale multicommodity routing. *Computational Optimization and Applications*, 27(2):187–215, 2004.

[79] J. Li and S. Sampalli. QoS-guaranteed wireless packet scheduling for mixed services in HSDPA. In *Proc. of the 9th ACM Intl. Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWiM 2006)*, pages 126–129, Oct. 2006.

[80] X. Lin, N. B. Shroff, and R. Srikant. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected Areas in Communications*, 24(8):1452–1463, Aug. 2006.

[81] J. Linderoth and M. W. P. Savelsbergh. A computational study of search strategies for mixed integer programming. *INFORMS Journal on Computing*, 11(2):173–187, Feb. 1999.

[82] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *Journal of the ACM*, 41(5):960–981, 1994.

[83] H. Marchand, A. Martin, R. Weismantel, and L. Wolsey. Cutting planes in integer and mixed integer programming. *Discrete Applied Mathematics*, 123(1):397–446, Nov. 2002.

[84] S. Martello and P. Toth. *Knapsack problems: Algorithms and Computer Implementations*. Wiley, Chichester, 1990.

[85] R. Mathar and T. Niessen. Optimum positioning of base stations for cellular radio networks. *Wireless Networks*, 6(6):421–428, 2000.

[86] F. F. Mazzini, G. R. Mateus, and J. M. Smith. Lagrangean based methods for solving large-scale cellular network design problems. *Wireless Networks*, 9:659–672, 2003.

[87] L. Mendo and J. M. Hernando. On dimension reduction for the power control problem. *IEEE Transactions on Communications*, 49(2):243–248, Feb. 2001.

[88] J. E. Mitchell. *Handbook of Applied Optimization*, chapter Branch-and-Cut Algorithms for Combinatorial Optimization Problems, pages 65–77. Oxford University Press, 2002.

[89] J. Mitola. The software radio architecture. *IEEE Communications Magazine*, 33(5):26–38, May 1995.

[90] K. Mnif, B. Rong, and M. Kadoch. Virtual backbone based on MCDS for topology control in wireless ad hoc networks. In *Proc. of the 2nd ACM Intl. Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks*, pages 230–233, 2005.

[91] A. Muqattash, T. Shu, and M. Krunz. On the performance of joint rate/power control with adaptive modulation in wireless CDMA networks. In *Proc. of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2006)*, pages 1–12, Apr. 2006.

[92] M. Naghshineh and I. Katzela. Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. *IEEE Personal Communications*, 3:10–31, June 1996.

[93] M. Nawrocki, H. Aghvami, and M. Dohler, editors. *Understanding UMTS Radio Network Modelling, Planning and Automated Optimisation: Theory and Practice*. John Wiley & Sons, Apr. 2006.

[94] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley, 1988.

[95] S. Rajasekaran. *Encyclopedia of Optimization*, chapter Randomization in Discrete Optimization: Annealing Algorithms. Oxford University Press, 2001.

[96] J. Razavilar, K. J. R. Liu, and S. I. Marcus. Optimal rate control in wireless networks with fading channels. In *Proc. of the 49th IEEE Vehicular Technology Conference (VTC '99)*, pages 807–811, 1999.

[97] C. R. Reeves and J. E. Rowe. *Genetic Algorithms, Principles and Perspectives: A Guide to GA Theory*, volume 20 of *Operations Research/Computer Science Interfaces Series*. Kluwer, 2002.

[98] I. Rhee, A. Warrier, J. Min, and L. Xu. DRAND: Distributed randomized TDMA scheduling for wireless ad-hoc networks. In *Proc. of the seventh ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '06)*, pages 190–201. ACM Press, May 2006.

[99] C. Roos, T. Terlaky, and J.-P. Vial. *Theory and Algorithms for Linear Optimization: An Interior point Approach*. Wiley, 1997.

[100] T. Salonidis and L. Tassiulas. Asynchronous TDMA ad hoc networks: Scheduling and performance. *European Transactions on Telecommunications*, 15(4):391–403, 2004.

[101] Y. Shang and S. Cheng. *Networking and Mobile Computing*, volume 3619 of *Lecture Notes in Computer Science*, chapter An Enhanced Packet Scheduling Algorithm for QoS Support in IEEE 802.16 Wireless Network, pages 652–661. Springer, Berlin, 2005.

[102] H. D. Sherali, M. Pendyala, and T. Rappaport. Optimal location of transmitters for micro-cellular radio communication system design. *IEEE Journal on Selected Areas in Communications*, 14(4):662–673, 1996.

[103] G. Sierksma. *Integer and Linear Programming: Theory and Practice*. Marcel Dekker, 1996.

[104] IST-2000-28088 MOMENTUM. `http://momentum.zib.de`, 2003 (updated in 2005).

[105] R. N. Tomastik, P.B. Luh, and D. Zhang. A reduced-complexity bundle method for maximizing concave nonsmooth functions. In *Proc. of the 35th IEEE Conference on Decision and Control*, pages 2114–2119, Dec. 1996.

[106] K. Tutschku. Demand-based radio network planning of cellular mobile communication systems. In *Proc. of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '98)*, pages 1054–1061, Apr. 1998.

[107] R. L. M. J. van de Leensel, C. P. M. van Hoesel, and J. J. van de Klundert. Lifting valid inequalities for the precedence constrained knapsack problem. *Mathematical Programming*, 86(1):1436–4646, Sep. 1999.

[108] F. Vanderbeck. On Dantzig-Wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm. *Operations Research*, 48(1):111–128, Jan.-Feb. 2000.

[109] R. J. Vanderbei. *Linear Programming: Foundations and Extensions*. Kluwer, Boston, 1996.

[110] V. Černy. A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985.

[111] Y. Wu, P. A. Chou, and S.-Y. Kung. Minimum-energy multicast in mobile ad hoc networks using network coding. In *Proc. of IEEE Information Theory Workshop*, pages 304–309, Oct. 2004.

[112] X. Zhao, P. B. Luh, and J. Wang. The surrogate gradient algorithm for Lagrangian relaxation method. In *Proc. of the 36th IEEE Conference on Decision and Control*, pages 305–310, 1997.

Part I

Network Planning and Resource Optimization for UMTS

Chapter 2

Introduction to UMTS Networks and CPICH Power Management

The high price of 3G licences and the growing competition in telecommunications market are putting enormous pressure on manufacturers and operators to configure the next-generation mobile network in the most cost-effective way possible providing high quality 3G services. Therefore, aiming to improve overall network performance and to ensure efficient utilization of network resources, radio network planning and optimization became even more important than in previous generation networks. On the other hand, in 3G networks, due to support of multiple services, high data rates, low frequency reuse, higher spectral loading and interference levels, and much uncertainty about future traffic growth, these processes became more challenging tasks. Coverage planning and optimization of radio base station antenna configuration and control channel power are among the issues the operators have to deal with. These optimization problems are addressed in the current part of the thesis.

2.1 3G Networks

2.1.1 The Evolution towards 3G

The idea of using cells for communication can be traced back to Bell Laboratories in 1947, but it was not until 1979 that such a system was first deployed in Japan. The United States followed with a system in Chicago in 1983. First generation (1G) is referred to as an analog technology since the radio frequency carrier is modulated using frequency modulation. By today's standard, 1G is archaic. It suffers from poor voice quality, poor battery life, large phone size, no security, frequent call drops, limited capacity, and poor handover reliability between cells. However, despite its limitations, 1G was a huge success, and 1G networks are still operational in many countries.

Through the use of digital technology, e.g., digital vocoders, forward error correction, high-level digital modulation, and greater use of computer technology, second generation (2G) networks provided improvements to system capacity, security, performance, and voice quality. Compared to 1G networks which use Frequency Division Multiple Access (FDMA) to support multiple users simultaneously, 2G networks use more advanced multiple access technologies such as Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) that allow for more efficient use of frequency spectrum. While 1G FDMA provides multiple access by separating users by radio frequency, TDMA achieves the goal by allocating time slots among users, and CDMA allows users to simultaneously access the available bandwidth by utilizing a coding scheme and separating users by codes. TDMA-based Global System for Mobile communications (GSM) was the first digital wireless technology, and the most popular, counting nowadays over two billions subscribers worldwide [24]. Another well-known 2G representative is cdmaOne (IS-95A), a CDMA-based technology widely spread in North

and South America and parts of Asia.

Customer demand for digital services is the major impetus for the third generation (3G) networks. However, due to a huge technical jump from 2G to 3G, 2.5G was proposed as a "bridge" technology that allows service providers to smoothly move from 2G to 3G systems and to provide customers limited 3G features before 3G is fully available. 2.5G systems use improved digital radio and packet-based technology with new modulation techniques to increase data rates, system efficiency, and overall performance. Among the other advantages are compatibility with 2G systems, possibility of a low-cost move towards 3G, and transparent to users transition from 2G to 3G. General Packet Radio System (GPRS) built on GSM technology and cdmaOne (IS-95B) built on cdmaOne (IS-95A) represent 2.5G. Enhanced Data rates for GSM Evolution (EDGE) is a more advanced standard than those of 2.5G, but it still does not meet all the requirements for a 3G system, e.g., speeds of up to 2 Mbps. Nevertheless, EDGE is more often referred to as a 3G system than to as a 2.5G system.

3G is based on an International Telecommunication Union (ITU) initiative for a single global wireless standard called International Mobile Telecommunications-2000 (IMT-2000) [33, 32]. This concept of a single standard evolved into a family of five 3G wireless standards which were approved in May 2001. The 3G radio access standards, together with the underlying technologies stated in square brackets, are shown below.

- IMT-DS (Direct Spread) [WCDMA]
- IMT-MC (Multi-Carrier) [CDMA2000, including 1X, 1XEV, and 3X]
- IMT-TC (Time-Code) [UTRA TDD, TD-SCDMA]
- IMT-SC (Single Carrier) [UWC-136/EDGE]
- IMT-FT (Frequency-Time) [DECT]

IMT-DS and IMT-MC are the 3G CDMA standards and the successors to GSM and cdmaOne, respectively. IMT-TC is a 3G standard based on a combination of TDMA and CDMA, IMT-SC is a 3G TDMA standard, and IMT-FT is a 3G standard that combines the features of TDMA and FDMA. Of these five standards, only first three allow full network coverage over macro cells, micro cells and pico cells, meet all the 3G requirements, and can thus be considered as full 3G solutions. EDGE, as it has been mentioned previously, cannot be considered as a full 3G solution so far. The last standard, IMT-FT, was defined by ETSI and it is well-known as Digital Enhanced Cordless Telecommunications (DECT). DECT is used for cordless telephony and could be used for 3G short-range "hot spots". Hence, it could be considered as being a part of a 3G network, but it cannot give full network coverage.

According to ITU and IMT-2000, a 3G standard must meet the following minimum requirements:

→ High-speed data transmissions; 3G data rates fall into three categories:
  · 2 Mbps in fixed or indoor environments,
  · 384 Kbps in pedestrian or urban environments,
  · 144 Kbps in wide area mobile environments,
  · Variable data rates in large geographic area systems (satellite);
→ Greater (compared to 2G) capacity;
→ Symmetric and asymmetric data transmission support;
→ Global roaming across networks and compatibility with 2G networks;
→ Improved security;
→ Enabling rich data applications such as VoIP, video telephony, mobile multimedia, interactive gaming and more;
→ Improved voice quality, i.e., comparable to that of wire-line telephony;
→ Support of multiple simultaneous services.

More advanced and higher capability, compared to previous generation, 3G technologies have become very attractive for mobile telecommunications business due to increasing traf-

fic demand, possibility of providing new services and more functionality as well as higher quality requirements set by users and growing competition from other wireless technologies. However, this business sector was not growing very fast in the beginning due to high prices of auction-based 3G licences and the growing competition. Thus, the first 3G network (based on WCDMA) was launched in October 2001 in Japan by NTT DoCoMo, but it was not until 2003 when 3G networks were also launched in other countries. Today the number of WCDMA subscribers, including HSPA, is more than 37 million in Japan and exceeds 115 millions over the world (as of July 2007 [24]). The number of all reported 3G CDMA subscribers is over 461 million, including CDMA2000, WCDMA, and their evolutions. By the number of launched 3G commercial networks, WCDMA/HSPA technology, being used in 164 networks launched in 73 countries out of 239 networks, leads the market with 68 % market share (as of July 2007 [24]).

The work presented in the current part of the thesis focuses on Universal Mobile Telecommunications Service (UMTS) networks that adopt WCDMA technology which is discussed in the next section.

2.1.2 Wideband Code Division Multiple Access (WCDMA)

Wideband Code Division Multiple Access (Wideband CDMA, or WCDMA) was developed by NTT DoCoMo as a radio interface for their 3G network FOMA (Freedom Of Multimedia Access) and later accepted by ITU as a part of the IMT-2000 family of 3G standards. It is one of the two main 3G technologies implemented in nowadays 3G networks.

WCDMA is a wideband spread-spectrum radio technology that utilizes the direct sequence CDMA signalling method to achieve higher speeds and support more users compared to TDMA-based 2G GSM networks. WCDMA can operate in two modes: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). WCDMA FDD uses paired 3G spectrum: two separate 5 MHz carrier frequencies are allocated for the uplink and downlink, respectively. In the TDD mode, 5 MHz is time-shared between the uplink and downlink. The TDD mode enables more efficient use of the spectrum for asymmetric services, e.g., Internet applications, whilst FDD was originally designed for symmetric services. More information on WCDMA FDD and TDD can be found in [12, 18, 27].

In WCDMA, transmitting and receiving users utilize the whole available frequency band. The transmitted data is encoded using a spreading code specific to a certain user such that only the intended receiver is able to decode the signal. To the others, the signal will appear as noise. The spreading codes used in WCDMA are Orthogonal Variable Spreading Factor (OVSF) codes and they must remain synchronous to operate. Because exact synchronization cannot be achieved in multipath environment, scrambling codes (pseudo random noise codes) are used to identify the individual transmissions. Thus, there are two stages of spreading. The first is applying an OVSF code and the second is using a scrambling code. At the receiver, the original signal is extracted by exactly the same spreading code sequence. This technique allows for simultaneous user transmissions over the entire spectrum. Furthermore, WCDMA has a frequency reuse of one, meaning that all users have the same frequency band to transmit the information. Among the other WCDMA features are inter-cell asynchronous operation, variable rate transmission, adaptive power control, downlink transmit diversity, supported by the standard multiuser detection and smart antennas.

WCDMA was selected as the radio interface for UMTS. The radio access for UMTS is known as Universal Terrestrial Radio Access (UTRA). UTRA operates in two modes: FDD, which adopts WCDMA FDD and is implemented in UMTS networks, and TDD, which is a part of UMTS-TDD networks and is based on TD-CDMA (Time Division CDMA) technology representing IMT-TC, another 3G standard. (UMTS and UMTS-TDD are thus not directly compatible.) Moreover, UMTS-TDD is typically not viewed as a full 3G solution and is mainly used to provide high data rate coverage for traffic hot spots and indoor environments.

The specification of UTRA has been created within the Third Generation Partnership

Project (3GPP), a joint standardization project of the standardization bodies from Europe, Japan, Korea, the USA, and China. UTRA has been formalized in several releases of 3GPP specifications. 3GPP Release 99 (March 2000), called "Major RAN release", specifies the architecture and the main UTRA aspects. Release 2000 was broken into two parts, Release 4 and Release 5, that together aim to create an all-IP network. Release 4 (March 2001), "Minor release", specifies some changes to the radio access with QoS enhancements. Release 5 (March 2002) includes HSDPA and IP Multimedia Subsystem (IMS) specifications. Release 6 (Dec. 2004) introduced integrated operation with WLAN and added, among the others, HSUPA.

Further in the thesis we do not take into account the UTRAN enhancements introduced by 3GPP in Release 5 and beyond, e.g., HSDPA and HSUPA. However, since HSDPA and HSUPA are deployed on top of the conventional WCDMA (either on the same or another carrier) and can share with it all the network elements in the radio and the core networks, the architecture model presented in the next section and the optimization approaches presented further in the current part of the thesis are also applicable to the evolved WCDMA networks.

2.1.3 UMTS Network Architecture

As can be seen from Figure 2.1, the architecture model for a UMTS network is composed of three parts [1],

- core network,
- UMTS terrestrial radio access network,
- user equipment.

The *core network* (CN) consists of physical entities that provide support for the network features and telecommunication services. The support includes functionality such as the management of user location information, control of network features and services, the transfer (switching and transmission) mechanisms for signalling and for user generated information. The core network enables communication between the depicted radio network and other networks, e.g., other public land mobile networks (PLMNs), Intranet, or any other external

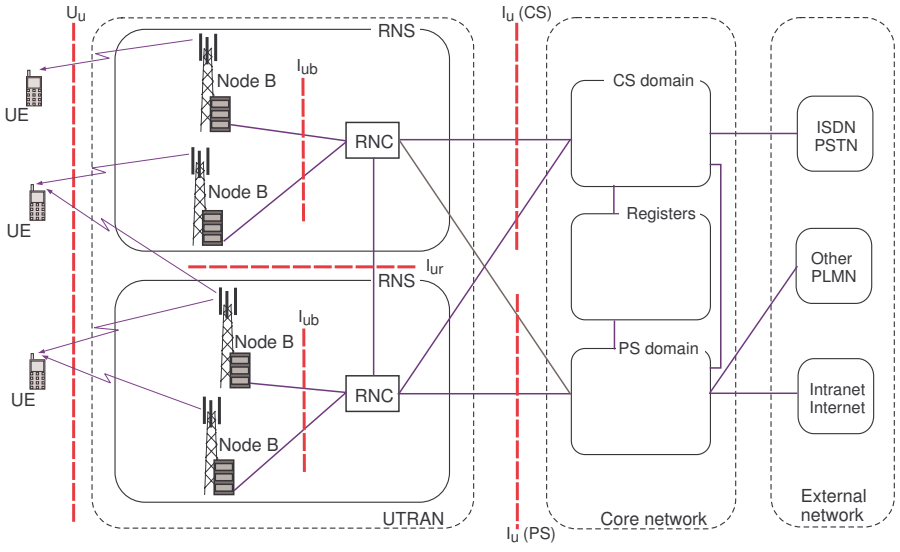


Figure 2.1: UMTS network architecture.

networks. The external networks can be divided into two groups: circuit switched networks, which provide circuit switched connections, and packet switched networks, which provide connections for packet data services. Integrated services digital network (ISDN) and public switched telephone network (PSTN) are examples of circuit switched networks. Internet is an example of a packet switched network.

In UMTS networks, *user equipment* (UE) is the equipment used by the user to access UMTS services. A UE consists of the *mobile equipment* (ME) and the *UMTS subscriber identity module* (USIM). The ME may be further subdivided into several entities, e.g., the one which performs the radio transmission and related functions, or *mobile termination*, and the one which contains the end-to-end application, or *terminal equipment*. The USIM smartcard and the ME are connected through the electrical interface that follows a standard format for smartcards.

The part of the network, which is conceptually located between the UE and CN, is commonly referred to as a *radio access network* and it implements a particular radio access technology. In UMTS, this part of the network is known as *UMTS terrestrial radio access network* (UTRAN) [2] and it implements the WCDMA radio technology discussed in Section 2.1.2.

The UE is connected to the UTRAN through the U_u interface, whereas the CN is connected to the UTRAN through the I_u interface. The I_u interface consists of two parts, the packet switched, $I_u(PS)$, and the circuit switched, $I_u(CS)$. Each separate I_u part connects the UTRAN with the corresponding core network domain, i.e., either the packet switched domain or the circuit switched domain.

The UTRAN is divided into *radio network subsystems* (RNSs). One RNS consists of radio elements and their controlling element. The radio elements are *radio base stations* (RBSs). A commonly used name for UMTS RBS is *Node B*, which was initially adopted by 3GPP as a temporary term during the specification process, but then never changed. The physical components of an RBS compose an RBS *site*. The controlling element is *radio network controller* (RNC). Each RNC is linked with a number of RBSs via the I_{ub} interface. Moreover, the RNCs are also interconnected with each other via the I_{ur} interface, which is needed to deal with UE mobility (e.g., soft handover).

From the point of view of the radio network and its control, each RBS consists of several entities called *cells*. A cell is the smallest radio network entity having its own identification number (cell ID), which is publicly visible for the UE. Every cell has one scrambling code, and a UE recognizes a cell by two values, scrambling code (when logging into a cell) and cell ID (for radio network topology). One cell may have one or several transmitter-receivers (TRXs, also called *carriers*) under it. The term *sector* stands for the physical occurrence of the cell, i.e., radio coverage.

In a UMTS network, the system has a hierarchical, or multi-layer, cell structure that consists of pico, micro, macro, and satellite cells. In pico cells, with low user mobility and smaller delay spread, high bit rates and high traffic density can be supported with low complexity. In larger macro cells, only low bit rates and traffic load can be supported because of the higher user mobility and higher delay spread. Global/satellite cells can also be used to provide area coverage where macro cell constellations are not economical to deploy or support long distance traffic. To separate cell layers within the same coverage area, different carrier frequencies are used.

2.2 Planning and Optimization for UMTS Networks

This section presents the major challenges in UMTS networks planning and optimization, and describes the planning process as well as the major issues arising in each phase of the process.

2.2.1 Challenges Arising in UMTS Networks

With the introduction of WCDMA-based networks, to which UMTS networks also belong, new challenges of radio network planning came up. These challenges are forced by two main aspects. On one hand, WCDMA has been developed mainly to support new data services with higher and variable data rates. On the other hand, the particular aspects of the underlying WCDMA radio access method impose fundamental changes in the planning methodology. The key properties of WCDMA are presented below.

Soft/softer handover

Soft handover is a feature specific to CDMA systems. User equipment and RBSs use special rake receivers that allow each UE to simultaneously communicate with multiple RBSs. The diversity gain associated with soft handover allows to improve the network performance.

Power control

Transmissions by the UE must be carefully controlled so that all transmissions are received with roughly the same power at the base station. If power control is not used, a "near-far" problem, where mobiles close to the base station over-power signals from mobiles farther away, occurs. The base station uses a fast power control system to direct the mobile to power up or power down as its received signal level varies due to changes in the propagation environment. Likewise, on the downlink, transmissions from the base stations are power-controlled to minimize the overall interference throughout the system and to ensure a good received signal by the UE.

Frequency reuse of one

In general, every RBS in a WCDMA-based radio access network operates on the same frequency for a given carrier and also all UEs share a common frequency within the network, i.e., the frequency reuse factor is one in such networks. Therefore, no frequency planning is required. However, since every cell causes interference to every other cell and every mobile interferes with any other, attention must be paid to controlling interference which typically depends to a large extent on inter-site distance, RBS antenna configuration, the amount of served traffic, and the cell coverage area.

Hierarchical cell structure

Although in general the frequency reuse factor is one for WCDMA, it doesn't mean that frequency reuse, the common CDMA property, cannot be utilized at all in UMTS networks. It is utilized by organizing the multi-layer cell structure consisting of pico, micro, macro and satellite cells. This shows the need of more careful control of inter-frequency handover. However, due to typically better space separation of cells, for example, at micro and pico layers used for indoor and hotspot areas, frequency planning in UMTS networks is less crucial than cell planning with a focus on interference management.

Soft capacity

Capacity and coverage are intertwined in CDMA, depending on the number of users and the amount of traffic in the system as well as on the amount of interference allowed before access is blocked for new users. By setting the allowed interference threshold lower, coverage will improve at the expense of capacity. By setting the threshold higher, capacity will increase at the expense of coverage. Because of the fundamental link between coverage and capacity, cells with light traffic loads inherently share some of their latent capacity with more highly loaded surrounding cells. The maximum capacity is thus limited by the amount of interference in

the air interface (soft capacity) and depends a lot on the spatial distribution of users within a cell.

Cell breathing

In CDMA systems the coverage of the cell expands and shrinks depending on the number of users. This is known as cell breathing and occurs because with CDMA, users transmit at the same time (and are identified by their unique code). The effect of cell breathing is based on the cell dominance concept and it is more complicated in WCDMA due to a frequency reuse of one.

The problem of planning second-generation cellular TDMA-based systems has usually been simplified by subdividing it into a coverage planning problem and a frequency planning problem. The corresponding radio base station placement and the frequency assignment problems have been discussed in Section 1.1.2. In the coverage planning phase, base stations are placed so that the signal strength is high enough in the area to be served [41, 52]. The coverage area achieved by a single antenna depends mainly on the propagation conditions and is independent from all other antennas in the network. As a result, in the frequency planning phase, a set of channels has to be assigned to each base station, taking into account the traffic requirements and the service quality measured as the signal-to-interference ratio.

In contrast, the network planning methodology for UMTS networks is much more complicated. The cell coverage area in a CDMA system does not only depend on propagation conditions, but also on the traffic load of the cell. The planning task becomes even more complicated due to mixed traffic scenarios (packet and circuit switched), mixed services with different bit rates, etc. Furthermore, the amount of interference received from other cells depends on their traffic load as well. Additionally the traffic load of a cell is influenced by the soft and softer handover areas. Thus, coverage, network configuration parameters, and capacity planning cannot be considered independently of each other, which is the main difference between the network planning for UMTS and GSM networks. With increasing complexity of the network planning and optimization tasks, automatization of these processes has become inevitable.

2.2.2 Automated Network Planning and Optimization

Automated radio network planning and resource optimization for UMTS networks are crucial engineering tasks that have attracted an increasing interest during the last several years (see, for example, [13, 25, 26, 38, 44, 48]). Automatization of these two processes allows operators to deal with the UMTS network design complexity, which is often beyond the reach of a manual approach. In addition to making the network design process time-efficient, planning tools incorporating automated optimization can significantly reduce network deployment and maintenance costs. UMTS network optimization involves a trade-off between many factors, such as service coverage, network capacity, quality of service (QoS), equipment costs, and expected revenues from network operation (see Figure 2.2). From a long-term perspective, the primary objective of an operator is to maximize the revenue. This objective plays a major role in the *network definition* phase. Issues involved in this phase include the choice of technology and its expected evolution, deployment strategy, service specification, as well as coverage and capacity requirements. Each phase of a network life cycle also involves short-term objectives and goals. For example, minimizing equipment cost is very important in *network dimensioning* when major equipment investments are required. Equipment cost is also of particular importance in *network expansion and upgrade*. In *detailed planning* as well as *network operation and maintenance*, the type and amount of equipment are typically given, and the focus is on optimizing network configuration and parameter setting to achieve the best possible network performance.

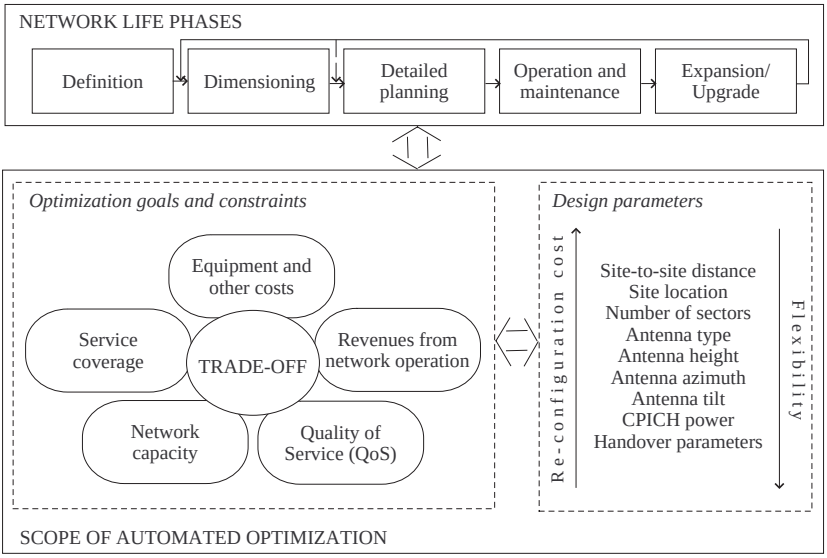


Figure 2.2: Network planning and optimization.

Various network planning phases involve different optimization problems. For example, in the network dimensioning phase, based on the operator's requirements (coverage, capacity, QoS requirements, etc.), evaluating possible configurations and the amount of network equipment needed is to be done [9, 41]. Next, using the predicted service demand, user distribution, and propagation measurements, the main tasks become site planning, detailed coverage and capacity planning, initial network configuration, and parameter tuning [13]. In the network maintenance phase, examples of optimization tasks are capacity maximization, load balancing, and improving the link budget by adjusting different performance parameters [26]. These problems are to be solved subject to different constraints such that the minimum required coverage degree, guaranteed QoS level for each customer group, limitation on availability of radio resources, etc..

The network planning and optimization tasks addressed in Chapters 3-5 are a part of automated optimization and detailed planning after network roll-out. In detailed planning, making major changes in network topology and layout is typically not an acceptable option for an operator. Instead, the goal is to optimize some key configuration parameters including antenna azimuth, antenna tilt, pilot power, and soft handover parameters. These parameters are at different levels of flexibility. For example, changing antenna azimuth requires higher effort and cost in comparison to electrical tilting. In the next section, there will be discussed in more detail three key RBS configuration parameters that affect network coverage: Common Pilot Channel (CPICH) transmit power, antenna tilt, and antenna azimuth. The CPICH power determines service coverage. To utilize power resource efficiently, CPICH should not be set to more than what is necessary to guarantee coverage. The optimal level of CPICH power, in its turn, depends on the other two parameters, i.e., the tilt and azimuth of RBS antennas. Therefore, when planning network coverage while optimizing radio resources, it is natural to optimize these three parameters jointly.

2.2.3 Radio Base Station Configuration Parameters

CPICH Transmit Power

In a UMTS network, a cell announces its presence through the CPICH, a fixed-rate downlink physical channel carrying a pre-defined bit/symbol sequence. Typically, each cell has one CPICH. CPICH signals, or *pilot signals*, are used by mobile terminals for channel quality estimation, cell selection/re-selection, and handover evaluation.

From a resource management standpoint, satisfying service coverage requirement using a minimum amount of CPICH power offers several performance advantages. Since the maximum transmit power available at RBS is constant, less power consumption by CPICH makes more power available to traffic channels. This benefit becomes particularly significant if the common practice of setting the power levels of some common channels relative to that of CPICH is adopted in the network [38]. Moreover, excessive pilot power adds to the total DL interference as well as increases cell overlap and potential pilot pollution area. On the other hand, coverage problem will arise if the CPICH power becomes too low. Since the entire Part I focuses on coverage control and pilot power optimization, more technical details and related issues will be discussed later in a separate section (Section 2.3).

Antenna Tilt

Antenna tilt is the angle of the main beam of an antenna below the horizontal plane. The primary goal of antenna downtilting in a UMTS network is to reduce the inter-cell interference in order to increase the relative strength of signals from the home cell. Interference reduction increases cell capacity and improves performance of the entire network. However, if antennas are downtilted excessively, the coverage may suffer. There are two ways of tilting an antenna — *mechanical* and *electrical*. The two tilting methods have different effects on signal propagation and therefore affect differently on the total DL interference [39].

Mechanical tilting means to adjust the physical angle of the brackets in which an antenna is mounted. Electrical tilt does not change the physical angle of an antenna, but adjusts the radiating currents in the antenna elements to lower the beam in all horizontal directions. This also changes the antenna characteristic which is not affected if the antenna is downtilted in a mechanical way. Note also that interference radiation of an electrically downtilted antenna is smaller compared to a similar mechanical downtilt. Furthermore, the coverage area is more affected when electrical tilt is applied. Therefore, in a coverage-limited environment (e.g., in rural areas), mechanical tilting can be more useful. In a capacity-limited environment (e.g., in a city center), minimization of the interference is more vital and hence electrical downtilt can provide better performance [39].

The radio propagation effect of an antenna in different directions can be derived from its horizontal and vertical diagrams, usually provided by manufactures, that show the antenna gain in all directions relative to an isotropic antenna. Examples of a horizontal and a vertical diagrams in polar coordinates for a directional antenna are demonstrated in Figure 2.3. Figure 2.4 shows the difference in antenna gains in 3D space obtained by interpolation [13] for an antenna downtilted by $\phi = 6^\circ$ using mechanical and electrical tilt.

If the antenna specification has an option of electrical tilting, electrical and mechanical tilt can be combined. Recently, the procedure of electrical downtilting has been significantly simplified by introducing remote electrical tilt (RET) controllers, which allows the operators to avoid costly site visits by technical personnel. The resulting flexibility is one of the advantages of improving network performance via antenna downtilt. On the other hand, due to Electro-Magnetic Compatibility (EMC) regulations in many countries, the range of possible tilt values can be quite limited in practice, especially in city environments.

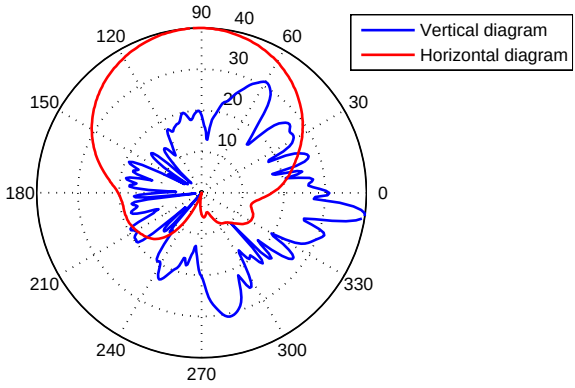
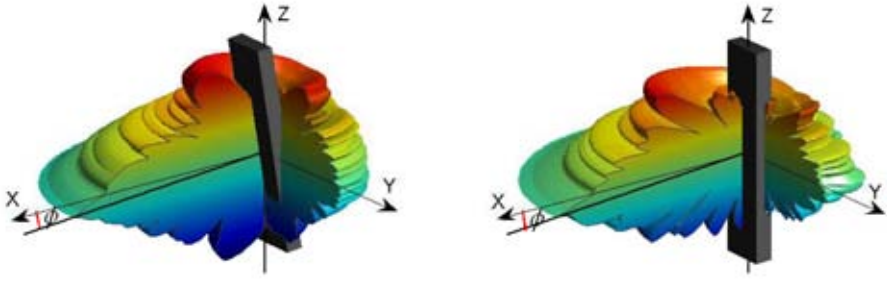


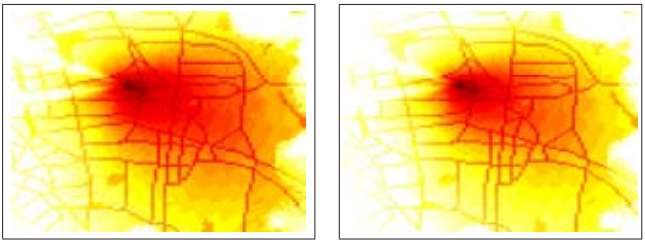
Figure 2.3: Examples of horizontal and vertical antenna diagrams.



(a) Mechanical downtilting by 6°

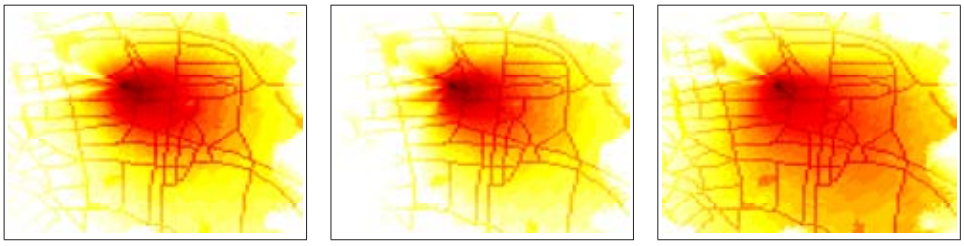
(b) Electrical downtilting by 6°

Figure 2.4: 3D interpolation of antenna diagrams and the tilting effect on radio propagation.



(a) Reference (2 W, 0°, 0°, 120°)

(b) Lower CPICH power (1 W, 0°, 0°, 120°)



(c) Mechanical downtilt (2 W, 6°, 0°, 120°)

(d) Electrical downtilt (2 W, 0°, 6°, 120°)

(e) Changed azimuth (2 W, 0°, 0°, 140°)

Figure 2.5: The effect of changing design parameters on the received CPICH signal strength.

Antenna Azimuth

Antenna azimuth is the horizontal angle between the north and the antenna's main lobe direction. Antenna azimuth is another configuration parameter having large influence on service coverage as well as cell overlap. Sometimes, adjusting antenna azimuth can significantly reduce cell overlap without sacrificing coverage. Less cell overlap, in turn, improves inter- and intra-cell interference, power consumption, and capacity. There exist various tools that simplify the task of steering azimuth. In most situations, however, adjusting azimuth has to be done manually.

In Figure 2.5, we illustrate the effect of changing CPICH power and antenna configuration on cell coverage. The example represents a part of one of our test networks (see Net8 in Appendix A). In the figure, the CPICH signal strength is represented by color. The values in parentheses denote CPICH power, mechanical downtilt, electrical downtilt, and azimuth, respectively.

2.3 Pilot Power Management in UMTS Networks

Because of the potential downlink limitations in WCDMA-based networks and utilization of downlink pilots for cell synchronization and handover control, tuning of downlink-related radio network configuration parameters are of critical importance. The focus of Part I of the thesis is CPICH power optimization which is used not only as a control parameter in the presented models but also as a measure of the amount of interference in the network when configuring RBS antennas.

2.3.1 Common Pilot Channel

The Common Pilot Channel, or CPICH, is a fixed rate (30 Kbps) downlink physical channel that carries a continuously transmitted pre-defined bit/symbol sequence [6]. The CPICH is an unmodulated code channel, which is scrambled with the cell-specific primary scrambling code. The function of the CPICH is to aid the channel estimation at the terminal for the dedicated channel and to provide the channel estimation reference for the common channels when they are not associated with the dedicated channels or not involved in the adaptive antenna techniques. Figure 2.6 shows the frame structure of a CPICH. The CPICH uses the spreading factor of 256 (which is the number of chips per symbol), and there are 10 pilot symbols in one slot. This gives 2560 chips per slot and thus 38400 chips per radio frame of 10 ms.

There are two types of common pilot channels, the Primary Common Pilot Channel (Primary CPICH, or P-CPICH) and the Secondary Common Pilot Channel (Secondary CPICH,

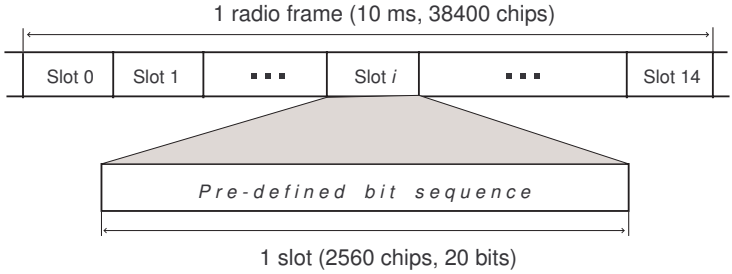


Figure 2.6: CPICH frame structure.

Table 2.1: Similarities and differences between P-CPICH and S-CPICH

Channel	Characteristics
Primary CPICH (P-CPICH)	• Always uses the same channelization code (code zero) • Scrambled by the primary scrambling code • Only one per cell • Broadcast over the entire cell • The primary CPICH is the phase reference for the SCH, primary CCPCH, AICH, PICH. It is also the default phase reference for all other downlink physical channels
Secondary CPICH (S-CPICH)	• Uses an arbitrary channelization code of spreading factor 256 • Scrambled either by the primary or a secondary scrambling code • Zero, one, or several per cell • Broadcast over entire or part of a cell • A secondary CPICH may be the reference for the secondary CCPCH and the downlink DPCH. If this is the case, the mobile station is informed about this by higher-layer signaling

or S-CPICH), both have the same frame structure shown in Figure 2.6. The difference is that the P-CPICH is always under the primary scrambling code with a fixed channelization code (code zero) allocation and there is only one such channel in a cell. Since there are 512 different primary scrambling codes available, typically it is possible to ensure that a single primary scrambling code is used only once in a particular area. S-CPICH may have any channelization code of length 256 and may be under a secondary scrambling code as well.

Normally, each cell has only one CPICH, P-CPICH, and it is used for broadcasting the pilot signal over the cell. In some cases, a cell may have also several additional CPICHs, S-CPICHs. Similar to P-CPICH, a secondary CPICH provides a coherent reference for demodulation of other channels, but it is typically not broadcasted over the entire cell and is mainly used to support coverage in dedicated hot-spot areas that are served by narrow beam antennas. In this case, a dedicated area uses the S-CPICH, whereas the P-CPICH broadcasts the pilot signal over the entire cell. This is because the CPICH intended as a phase reference for another channel and the channel itself must use the same antenna. The similarities and differences between P-CPICH and S-CPICH are summarized in Table 2.1.

The CPICH does not carry any higher layer information, neither is there any transport channel mapped to it. It may be sent from two antennas in case transmission diversity methods are used in the RBS. In this case, the transmissions from the two antennas are separated by a simple modulation pattern on the CPICH signal transmitted from the diversity antenna, called diversity CPICH. The diversity pilot is used with both open loop and closed loop transmit diversity schemes.

A mobile terminal makes two types of measurements on CPICH. The first is CPICH Received Signal Code Power (RSCP) which is used for handover evaluation, downlink (DL) open loop power control, uplink (UL) open loop power control, and for estimating the path loss (because the transmit power of the CPICH is either known or can be read from the system information). The second measurement is the CPICH E_c/I_0, the ratio of the received energy per chip for the CPICH to the total received power spectral density at the antenna connector of the mobile terminal. The ratio is also often referred to as *carrier-to-interference ratio*, or CIR. This measurement is used for cell selection/re-selection [7, 5] and for handover evaluation [5], and it is the most important mobile terminal measurement in WCDMA for the purpose of network planning since it typically has a good accuracy [38] and is used as the basic coverage indicator.

Cell selection is a procedure of choosing a cell for the mobile to camp on after a mobile terminal has switched on and has found a suitable network. The selection decision is made based on the measured E_c/I_0 values of the CPICH signals. Cell selection is preceded by *initial cell search* during which the mobile terminal tries to find a cell and determine the

cell's downlink scrambling code. The scrambling code information is needed to be able to receive system information.

Cell search is also performed when a mobile terminal is in the active or idle modes (*target cell search*). Target cell search is used to determine handover candidate cells and is triggered when the network requests the mobile terminal to report *detected set* cells. On this request the mobile terminal shall search for cells outside the monitored and active sets that together contain cells continuously measured by the terminal. Moreover, the RNC sends a set of thresholds that are used when the mobile terminal measures the cells and decides whether the measurements should be reported to the RNC. Based on the reported measurements, the RNC decides whether to update the list of cells to be measured or not. *Cell reselection* procedure is responsible for guaranteeing the required QoS by always keeping the mobile camped on a cell with good enough quality. Based on measurements of the monitored and active sets, the mobile terminal ranks the cells by the cell-ranking criterion and reselects the best cell if the reselection criteria are fulfilled during a certain time interval. The quality of cells is determined based on CPICH E_c/I_0 measurements.

Soft/softer handover is a function in which the mobile terminal is connected to several RBSs at the same time. The decisions about triggering soft/softer handover (SHO) are based on the comparison of the CPICH E_c/I_0 values between RBSs.

We have discussed how the CPICH measurements are used in different UTRAN procedures and have observed that the measurements play an important role for cell coverage. This also suggests that the CPICH measurements have a great impact on cell sizes and therefore can be utilized for controlling the load of a cell and load balancing among neighboring cells. Let us now examine the factors that define the CPICH quality. According to the definition of CPICH RSCP and CPICH E_c/I_0, the list of the factors is as follows,

- CPICH transmit power,
- attenuation between the antenna and user terminal,
- interference on the same channel (total received signal power from own and other cells),
- noise and adjacent channel interference.

Attenuation and the amount of received noise and adjacent channel interference depend on the environment and hardware. Downlink attenuation is also dependent on antenna configuration which, however, can be considered fixed in short-term planning. Interference on the same channel is mostly dependent on the amount of traffic since user traffic signals contribute to the total received signal power more than control channels. As a result, the CPICH transmit power can be viewed as an efficient and the only relatively autonomous factor for controlling the received CPICH signal strength, especially in highly loaded networks with constantly high interference. In a long run, antenna configuration can also be viewed as an effective source of improving the quality of the received CPICH signals due to interference reduction. Therefore, one of the extensions of the pilot power optimization problem studied in the thesis addresses joint optimization of CPICH power and RBS antenna configuration. Note, however, that finding an optimal (non-uniform) CPICH power setting in a network, even without considering antenna configuration, is not simple since this involves resolving a trade-off between power consumption and coverage. Challenges arising in pilot power control are discussed in the next section.

2.3.2 Pilot Power Control Challenges

Pilot power control in UMTS networks is a crucial engineering issue which has been attracting the attention of researchers in industry and academia during the last several years. In UMTS network, adjusting the P-CPICH power levels allows us to control cell sizes, the number of connected users to a cell and to balance the traffic among the neighboring cells, which in turn allows us to regulate the network load. The goal of pilot power control is to ensure that all RBSs use just enough transmit power to guarantee the required coverage and QoS in the

Table 2.2: Typical power allocation for the downlink common channels

Downlink common channel	Relative to CPICH	Activity	Average allocation with 20 W maximum power
Primary common pilot channel P-CPICH	0 dB	100 %	2.0 W
Primary synchronization channel SCH	-3 dB	10 %	0.1 W
Secondary synchronization channel SCH	-3 dB	10 %	0.1 W
Primary common control physical channel P-CCPCH	-5 dB	90 %	0.6 W
Secondary common control physical channel S-CCPCH	0 dB	10 %	0.2 W
Paging indicator channel PICH	-8 dB	100 %	0.3 W
Acquisition indicator channel AICH	-8 dB	100 %	0.3 W
Total amount of common channel power			∼3.6 W

entire network.

Pilot power control always involves a *trade-off between the pilot power consumption and coverage*. The transmit powers of the downlink common channels are determined by the network. In general, the relation between the transmit powers between different downlink common channels is not specified by the 3GPP standard and may even change dynamically. Usually, as a rule of thumb, for the P-CPICH a transmit power of about 30-33 dBm, or 5-10 % of the total cell transmit power capability, is allocated [27, 38]. In addition to P-CPICH, a cell uses a number of other common channels, and the transmit power of these channels is typically set in proportion to that of P-CPICH. An example of a power allocation setting [27] for the common channels is presented in Table 2.2.

The total downlink transmit power is shared between the control and traffic channels. Obviously, the more power is spent for control signalling, the less power is left to serve the user traffic. Excessive pilot power can easily take too large proportion of the total available transmit power, so that not enough power is left for traffic channels although more mobile terminals may be willing to handover to the cell if the cell size increases. Also, with introducing HSDPA, the amount of high-speed data traffic depends on the power left in the cell after allocating the necessary power to dedicated channel traffic. This makes efficient power allocation for common channels particularly important. Observe, for example, from Table 2.2 that although the amount of power allocated to other DL common channels does not exceed that of P-CPICH (is about 80 % of the P-CPICH power), the total amount of power consumed by all DL common channels, including P-CPICH, is almost 20 % of the maximum power of the cell. Reducing this amount to 10–15 % (which is feasible according to numerical studies presented in Chapters 3-5) would have a significant impact on network capacity. Decreasing the CPICH power leaves more power available for user traffic and therefore increases the cell capacity, but may also cause coverage problems. On the other hand, increasing the pilot power yields better coverage, but this is at a cost of less power available for user traffic.

There is another factor that needs to be considered in the trade-off between coverage and pilot power consumption – *soft and softer handovers*, which depends on coverage and affects power consumption and network performance. SHO improves the performance due to the macro diversity principle allowing to reduce the transmit powers both in uplink and downlink. On the other hand, in SHO, a mobile terminal is simultaneously linked to two or more cells which increases the traffic amount (SHO overhead). The SHO overhead must be kept within reasonable limits to save the traffic capacity of the cell, otherwise, the network capacity gain from macro diversity can vanish due to high SHO overhead or even turn into a network capacity loss.

Another goal of pilot power management is to control the amount of *interference* in the network. The strength of the CPICH signal affects the total interference in several ways. First, their transmit power adds to the total downlink interference in the network.

Second, higher CPICH transmit power increases cell overlap areas and the number of users in SHO. The latter may lead to higher interference in case SHO link gain is smaller than SHO overhead. Third, spacial distribution of both UL and DL interference depends on coverage areas of different cells that, in turn, depend on CPICH power.

CPICH signals provide cell–specific signals for RRM procedures, such as handover and cell selection/reselection. Detecting CPICH signals of approximately equal strengths or multiple strong CPICH signals at the same time may cause *pilot pollution*, i.e., pilot pollution can be observed in areas where a mobile terminal does not have enough RAKE fingers for processing all the received pilot signals or there is no dominant pilot signal at all [39, 40]. Pilot pollution cannot be totally avoided with traditional radio network planning methods due to inhomogeneous propagation environment and overlapping cells, but it can be reduced, for example, by optimizing the pilot powers in such a manner that required coverage thresholds are still exceeded [40]. The result is more clear cell dominance areas. Among the other instruments of reducing pilot pollution are optimizing antenna configurations [30, 46] and implementation of repeaters [11]. Repeaters can make the dominance area of a donor cell clearer, but, if not carefully planned, they can also shift pilot pollution interference and create pilot polluted areas in another location.

The use of CPICH reception level at the terminal for handover measurements has the consequence that by adjusting the CPICH power level, the cell load can be balanced between different cells, which also allows to control the load of hot spot cells, i.e. the cells that serve hot spot areas, and to improve network capacity. Reducing the CPICH power causes part of the terminals to hand over to other cells, while increasing it invites more terminals to handover to the cell as well as to make their initial access to the network in that cell. In WCDMA networks, load balancing is always associated with the effect of cell breathing which occurs when the cell load changes. The increasing cell load leads to the increased total received power which decreases the CIR and makes the cell shrinking. The dynamic cell breathing is the power management process, in which a base station automatically adjusts pilot power level as the cell load increases to ensure the balance between the uplink and the downlink handover boundary [57]. In some papers, this process is also referred to as technical cell breathing [47]. The problem of load balancing by adjusting pilot power levels has been addressed in [15, 22, 50, 54, 58, 60], but has not been studied in the thesis.

From the discussion in this section it follows that the pilot power assignment is a challenging task in planning and optimization for WCDMA networks. The problem of pilot power optimization is a multi-objective problem with many parameters. This kind of problem is very difficult to formulate and even more difficult to solve, especially for large real-life networks. Therefore, there is no standard approach for solving this problem. In most cases, pilot power assignment is based on a combination of professional experience and solutions to simplified optimization problems with different objectives. Section 2.3.3 gives a survey of the existing pilot power assignment approaches and related work.

2.3.3 Pilot Power Assignment Approaches and Related Work

Typically 5–10 % of the total downlink transmit power of the base station is used for P-CPICH [38], but there is no standard approach to find a pilot power setting. A number of existing approaches to resolve this issue are presented below. The most effective approaches are those based on optimization, but they are not always easy to implement.

In [38], Laiho et al. discuss several strategies for assigning the P-CPICH transmit powers of the individual cells, including the following,

- *Uniform P-CPICH power*. By this strategy, all cells use the same P-CPICH power.
- *Manual assignment*, by which the P-CPICH power levels are defined manually for each cell and can be different for different cells.

- *Maximum P-CPICH power for the lowest-loaded cell.* This approach is based on scaling the P-CPICH power levels, assigning the maximum P-CPICH power to the lowest-loaded cell (to make it more attractive) and scaling other cells' P-CPICH powers by the load relative to that cell.

Using uniform pilot power levels is the easiest and the most commonly used strategy in assigning the P-CPICH transmit power levels, although it is efficient only in simple propagation scenarios, where the signal attenuation is essentially determined by distance. In such scenarios, if fairly uniformly distributed traffic and equally-spread base stations are assumed, the sizes of the cells will be roughly the same. However, in an in-homogenous planning situation (e.g., a mix of rural and downtown areas), a uniform pilot power is not an efficient solution from the power consumption point of view. Moreover, such a solution also suffers from high total interference level in the network, big cell overlapping areas, and high pilot power pollution in many parts of the network [46]. The problem of minimizing the uniform pilot power for all (macro) cells has been addressed by Eisenblätter et al. in [14, 17], where pilot power optimization is a part of a more general network optimization problem, which the authors formulated as a MIP problem, that aims at optimizing site locations, base stations configuration, user assignment, and uplink and downlink transmit power assignment for traffic channels. The problem of optimizing the uniform pilot power is also covered in this thesis. In particular, optimal solutions that minimize the total amount of pilot power consumed in the network subject to a full coverage constraint and a partial coverage constraint are presented in Chapter 3 and Chapter 4, respectively. A simulated annealing algorithm for minimizing uniform CPICH power and adjusting antenna configuration parameters ensuring full coverage of the service area is presented in Chapter 5.

Mostly based on practical experience and professional intuition, manually defining the P-CPICH power level for each cell is usually more efficient than the uniform pilot power approach, but it often gives solutions that are far away from the optimal pilot power setting. Moreover, finding an optimal set of parameters for each cell manually is a tedious task, especially for large networks. Furthermore, operators often concentrate their efforts on troubleshooting work rather than time-consuming RAN optimization. This makes the manual approach to assigning the P-CPICH powers inappropriate and shows the need of techniques for pilot power management as well as for optimization of other network configuration parameters, which could be performed automatically. This would allow the operators to not only save man-hour resources but also benefit from more efficient planning. Thus, Love et al. in [40] demonstrated that a rule-based optimization technique for setting pilot power levels significantly outperforms a manually-designed solution in terms of network cost.

The pilot power assignment approach, by which the maximum P-CPICH power is set in the lowest-loaded cell, is an ad hoc approach which aims at balancing traffic load in the network. This approach is typically justified by experimental studies of the effect of adjusting pilot power on cell load and cell capacity. Such experiments have been conducted and the observations have been analyzed in a systematic way, for example, by Yang and Lin in [57], where the authors concluded that controlling the CPICH power level can be viewed as another means of load balancing and increasing capacity of heavily loaded cells. In [60], Zhu et al. studied by simulations load balancing by controlling CPICH power and proposed a set of key performance indicators (KPIs), among which are network throughput, DL load factor, DL call success rate, and DL bad quality call ratio, which have to be monitored when adjusting CPICH power. An approach for load balancing that ensures full coverage of the service area was presented in [50] where the authors considered the problem of maximizing the capacity ratio of the bottleneck cell. The cell capacity ratio was modeled as the amount of power in the cell available for traffic and the traffic power demand in the same cell.

In [53], Valkealahti et al. presented a cost-minimization approach that was implemented in a network simulation tool and studied by simulations. Based on target values for coverage and traffic load, the algorithm attempts to minimize the deviation from the target values by

adjusting the levels of the pilot power using a gradient-decent procedure. Solving a similar problem in [54], the authors presented a simple rule-based algorithm by which the CPICH power of a cell was increased or decreased by 0.5 dB if the cell load was significantly lower or higher, respectively, than the neighboring cells' load as indicated by the load statistic. Even if the load was not significantly unbalanced among the cells, but the CPICH signal reception was significantly lower or higher than the target, the pilot power was increased or decreased by 0.5 dB, respectively. In the simulator, the pilot power control was performed every second, assuming the CPICH power levels being limited by the minimum level of 3 % and the maximum level of 15 % of the maximum cell power. In [26], the CPICH power adjustment algorithm was presented as a built-in component of a higher-level approach to automated optimization of key WCDMA parameters.

A CPICH transmit power tuning algorithm for equalizing cell load while ensuring sufficient coverage in an irregular macro cellular system was proposed by Ying et al. in [58]. The algorithm consists of two parts running in parallel, "load sharing" and "coverage balancing". Load sharing is accomplished by tuning the CPICH transmit power individually per cell. In each iteration, the CPICH transmit power of a cell is adjusted based on the relative downlink traffic load of the cell. Coverage balancing is run as an outer loop mechanism to ensure a satisfactory CPICH outage probability level. Another approach for load balancing based on the simulated annealing optimization technique was presented by Garcia-Lozano et al. in [20], where the authors aimed at finding a CPICH power setting that minimizes the total UL transmit power and ensures the required coverage degree.

In [36], Kim et al. presented a problem of pilot power minimization subject to coverage constraints and a heuristic algorithm that adjusts the pilot power for one cell in each iteration. The disadvantage of the algorithm is that, due to the problem structure, it is very unlikely that it is able to find a globally optimal pilot power setting in the network but rather provides with a locally optimal solution. The optimal solution to this problem can only be obtained by optimizing the P-CPICH power levels for all cells simultaneously. This was a motivation for designing the models and optimization approaches presented in Chapters 3 and 4 of the thesis.

Since CPICH power and RBS antenna configuration are the parameters that affect both coverage and capacity, and they are typically used by operators for tuning performance of a deployed network, it is natural to optimize them jointly. This approach was followed, for example, by Gerdenitsch et al. in [22] where the authors aimed at maximizing network capacity using the grade-of-service measure, i.e., the portion of served users, as the objective. In each iteration, the presented algorithm changes antenna configuration parameters and CPICH power simultaneously based on a set of rules defined for the estimated network quality factor, which is a function of UL load, DL load, and OVSF utilization ratio. The authors reported a capacity increase of about 60 % in a network scenario with 75 cells.

The simulated annealing algorithm was used by Garcia-Lozano et al. in [21] to maximize the network capacity using as the objective function a capacity metric averaged over a number of snapshots. The optimization problem was subject to the constraints of the maximum allowed coverage loss and the maximum allowed portion of blocked users. To avoid non-feasible solutions, the authors applied a local search in each iteration. Simulated annealing was also used for multi-objective optimization by Picard et al. in [48] for the purpose of automatic cell planning. The considered objectives are maximization of capacity and coverage, and the decision is the optimal selection of the antenna model, antenna tilt and azimuth, and the transmit power of common channels.

In [43], Nawrocki et al. discussed the choice of the objective function for antenna optimization in UMTS networks and numerically analyzed the cost function features in a small network scenario for an objective function represented by the sum of cell load metrics. Niemelä and Lempiäinen studied by simulations the impact of RBS antenna downtilt on system capacity and network coverage in [45], where the authors derived an analytical

expression for the optimal geometrical downtilt angle as a function of the geometrical factor and antenna vertical beam width. Based on the simulated optimal downtilt angles (obtained in an ad hoc manner by changing them gradually), the authors also derived an empirical equation for computing antenna downtilt as a function of base station antenna height, the intended length of the sector dominance area, and the half-power vertical beam width. The simulated optimal downtilt angle was defined as a network-wide downtilt angle with which network coverage is guaranteed, and simultaneously, other-cell interference is mitigated as efficiently as possible.

To conclude, in practice, there exist plenty of approaches of tuning the network performance parameters. Many of them, however, are either not presented in the literature or are badly documented due to commercial interests involved. Many of the published solution approaches are simple rule-based algorithms that incorporate Monte-Carlo snapshot evaluations. Optimization-based approaches are used to less extent, although they have proved their ability to provide much better solutions. This is explained by computational time constraints, on one side, and by large problem size, typically difficult problem structure, integrality property of some parameters, and discontinuity of a feasible solution region, on the other side. Fast and reasonably good solutions, without solution quality estimates, are therefore often used in practice. To deal with this problem, optimization techniques have been also applied to reduce the problem size (see, e.g., [49]) and to derive compact structures that describe local network properties (like, for example, in [16, 34]). The goal of the UMTS part of the thesis is to develop mathematical models and optimization approaches suitable for optimizing configuration of large networks and study the results in realistic planning scenarios.

Chapter 3

A Basic Model for CPICH Coverage Optimization

3.1 System Model

Consider a UMTS network with a set of cells denoted by $\mathcal{I}$. The service area is represented by a grid of *bins* (small square or rectangular areas) with a certain resolution, assuming the same signal propagation conditions across every bin. The set of bins is denoted by $\mathcal{J}$.

Let P_i^{max} be the maximum available transmit power[1] available in cell i. The total amount of power available in the cell depends on the output power of the RBS power amplifier and the software parameters that define RRM in the cell. The output power of an RBS power amplifier is a hardware-specific parameter restricted by the 3GPP linearity requirements[2] [4]. For macro cells, the typical output power is 20–30 W. However, higher power amplifiers (e.g., 45 W) can be used, for example, in OTSR (Omni Transmit Sectorial Receive) configurations where the power of one or a few amplifiers is shared between several sectors of an RBS.

We use P_i^{Tot} ($P_i^{Tot} \leq P_i^{max}$) to denote the total allocated (actually used) DL transmit power in cell i. The power is used for both control signaling and traffic data in the cell. In a real network, the amount of power P_i^{Tot} depends on the current DL traffic. Moreover, the DL load of cell i is often measured as P_i^{Tot}/P_i^{max}. Therefore, total allocated DL power can be expressed as $P_i^{Tot} = \eta_i^{DL} \cdot P_i^{max}$, where η_i^{DL} is the *DL load factor*.

The amount of power allocated in cell i to CPICH is denoted by P_i^{CPICH} ($P_i^{CPICH} < P_i^{Tot}$). For a single cell, a higher value of P_i^{CPICH} means larger coverage area of cell i, but, on the other hand, less power available to serve user traffic in the cell and therefore a decrease in capacity. This is especially important if the transmit power levels of other common channels are set relative to CPICH, which is a common practice among operators [27]. Therefore, it is natural to require that the amount of CPICH power does not exceed some percentage (e.g., 10 % in [38]) of the total available power P_i^{max} in the cell.

We use a set of power gain parameters $\{g_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}$ to account for factors that affect the strength of the pilot signal received at a user terminal. Parameter g_{ij} ($0 < g_{ij} < 1$) is the power gain between the antenna in cell i and bin j. The parameter aggregates all losses and gains between the RBS transmit unit and the receiver of the user terminal. The power gain value depends on the antenna line characteristics (feeder cable, jumper, and connector losses, loss/gain due to mast head amplifier), antenna configuration (e.g., height, azimuth, mechanical tilt, electrical tilt), user's equipment, and path loss which depends on the radio propagation environment and distance. For a specific RBS equipment, the RBS and antenna characteristics are usually known from the manufacture or can be derived. Directional loss

[1] Linear scale is considered for all parameters and variables throughout the entire thesis unless other is mentioned.

[2] *Linearity* is the degree to which amplified signals remain within their prescribed band of the spectrum with low distortion or interference from adjacent channels.

for a given antenna configuration can be derived from antenna diagrams (see, for example, [13]). Therefore, the most critical aspect in computing the power gain parameters is the accuracy of path loss predictions.

Path loss predictions can be obtained by statistical (empirical), deterministic (theoretical), or hybrid models [44]. Statistical models are based on measurements in the environments of interest, whereas deterministic models apply the fundamental principles of radio wave propagation to a given environment. Statistical models are not as accurate as deterministic models that are able to better deal with the multi-path propagation environments. However, statistical models are computationally less expensive in general. The accuracy of statistical models can be improved by calibration, and the precision of propagation predictions can be controlled by tuning the model parameters. An example of statistical models used in macro cell planning is the COST 231 Hata model [37]. Among deterministic models, the most commonly used technique is based on the concept of artificial neural networks. The accuracy of the propagation predictions does not only depend on the chosen model, but also, among others, the accuracy in terrain information and data resolution (bin size). The latter is chosen to reach a trade-off between the amount of computational effort and the desired data accuracy, which depends on the planning task and the required level of details. For macro cells, the data resolution can well be 10 m and 100 m, depending on the intended use of data. Clearly, inaccuracy and insufficiently high precision level in the propagation data induce planning errors and may lead to wrong decisions. One way to account for possible prediction errors in coverage planning is to consider a higher target carrier-to-interference ratio (CIR) than a value typically used in the equipment.

By the 3GPP standard [5], to achieve the CPICH coverage of cell i in bin j, the E_c/I_0 of CPICH from cell i must meet a given threshold, i.e.,

$$\frac{g_{ij}\ P_i^{CPICH}}{I_j} \geq \gamma_0\ , \tag{3.1}$$

where $g_{ij}P_i^{CPICH}$ is the received CPICH power from cell i in bin j, I_j is the total received power spectral density in bin j, and γ_0 is the E_c/I_0 target. The E_c/I_0 threshold is specific for each user terminal, and it is fixed. The 3GPP standard [5] enforces the mobile terminals to be able to detect a pilot signal with $E_c/I_0 \geq -20$ dB. However, a slightly higher threshold, e.g., -18 dB, is usually used in practice (see, for example, [13, 44]).

We compute the amount of interference in bin j as

$$I_j = \sum_{l \in \mathcal{I}} g_{lj}\eta_l^{DL}P_l^{max} + \nu_j + I_j^A\ , \tag{3.2}$$

where ν_j is the thermal noise power in bin j, I_j^A is the adjacent channel interference in bin j. The adjacent channel interference is the amount of power leaking from an adjacent carrier either from the operator's own network or from the competitor's network due to non-linearity in power transmitters. Moreover, the user equipment is typically not able to completely filter out signals received on adjacent channels. For DL, the limit on the out-of-band power leakage is controlled by 3GPP [4] which specifies the Adjacent Channel Leakage power Ratio (ACLR) for RBS power amplifiers. The 3GPP standard [3] specifies the requirements on user equipment's ability to receive a signal at the assigned channel in the presence of an adjacent channel signal. The measure of this ability is known as Adjacent Channel Selectivity (ACS). In the own network, the operator can compute the adjacent channel interference as the total received power on the channel but with power gains scaled by the Adjacent Channel Protection (ACP) factor which is computed from ACLR and ACS. The information about the competitor's network is typically not available, therefore the adjacent channel interference in such a situation can be modeled as a grid of location-specific interference values obtained either empirically or theoretically.

Observe from (3.2) that the DL interference is at its maximum when $\eta_l^{DL} = 1.0$ for all cells. This scenario, also known as the *worst-case interference scenario*, is typically used for network modeling under the most pessimistic assumptions about the network load and is particularly relevant to planning networks with high traffic load or in uncertain traffic conditions, i.e., in an early network planning stage. In a stable network, the maximum DL load factor typically does not exceed 70–80 % [38]. The worst-case interference assumption becomes even more justified if coverage planning is done for a UMTS network enhanced with HSDPA where one of the power allocation strategies is to prioritize serving conventional dedicated channel traffic users and allocating the rest of available power to the high-speed traffic [28].

In order to utilize power resources efficiently, the CPICH transmit power should not be set to more than what is necessary. That is, to provide CPICH coverage in bin j, cell i does not need to use CPICH transmit power higher than P_{ij}, which can be derived from (3.1) as follows,

$$P_{ij} = \frac{\gamma_0}{g_{ij}} \cdot \left(\sum_{l \in \mathcal{I}} g_{lj} \eta_l^{DL} P_l^{max} + \nu_j + I_j^A \right) . \tag{3.3}$$

To achieve CPICH coverage, the received pilot signal must not only satisfy the CIR condition but also the minimum CPICH received signal code power (RSCP) requirement that enables the user equipment to properly decode pilot signals, i.e.,

$$g_{ij} P_i^{CPICH} \geq \gamma_1 , \tag{3.4}$$

where γ_1 is the receive sensitivity threshold. A typical value of γ_1 is in the range $[-120, -114]$ dBm. Thus, combining the CPICH E_c/I_0 and RSCP requirements, the minimum CPICH power needed to provide coverage by cell i in bin j can be found as follows,

$$P_{ij} = \max \left\{ \frac{\gamma_1}{g_{ij}}, \frac{\gamma_0}{g_{ij}} \cdot \left(\sum_{l \in \mathcal{I}} g_{lj} \eta_l^{DL} P_l^{max} + \nu_j + I_j^A \right) \right\} . \tag{3.5}$$

Our CPICH coverage modeling approach is demonstrated in Figure 3.1 where the CPICH RSCP is at least γ_1 in the area bounded by a thick solid line and the CPICH E_c/I_0 satisfies (3.1) in the colored area. Thus, the CPICH coverage in the this example is represented by the intersection of the two areas. In practice, the area with the CPICH signal satisfying the RSCP requirement is typically at least as large as the area where the CPICH E_c/I_0 threshold is met. This is especially true in urban scenarios where the interference is high. Therefore, planning for the CPICH E_c/I_0 coverage with the following up CPICH power adjustment with respect to the RSCP requirement is justified in practice.

Let Π_i^{max} ($\Pi_i^{max} < P_i^{max}$) be the upper limit for pilot power in cell i, and Π_i^{min} ($0 \leq \Pi_i^{min} \leq \Pi_i^{max}$) be the lower pilot power limit in cell i. The pilot power is said to be *unbounded*, if $\Pi_i^{min} = 0$ and $\Pi_i^{max} = P_i^{max}$, but such a scenario is unlikely to be used in a real-life network and therefore should also be avoided when planning a network. In WCDMA networks, the lower and the upper pilot power limits in a cell are usually set to 1–3 % and 15 % of the total transmission power of the cell, respectively [31, 53, 54].

The pilot power limits can be either introduced into the model as a set of constraints or they can be accounted for in the preprocessing step. The latter allows us to not only reduce the number of constraints, but also to significantly reduce the problem size. The smaller are the Π_i^{max} values, the smaller the sets of feasible pilot power levels become. To ensure the lower pilot power limit in the preprocessing step, P_{ij} values that are below Π_i^{min} are set to this minimum value. To consider the upper limits, P_{ij} values that exceed the upper limit Π_i^{max} are excluded from the list of possible pilot power settings. This can be done through the use of sets $\mathcal{I}_j$ and $\mathcal{J}_i$ which we define next.

For each bin j, we introduce a set $\mathcal{I}_j \subseteq \mathcal{I}$ which contains all the cells that may cover bin j with a feasible pilot power level, i.e.,

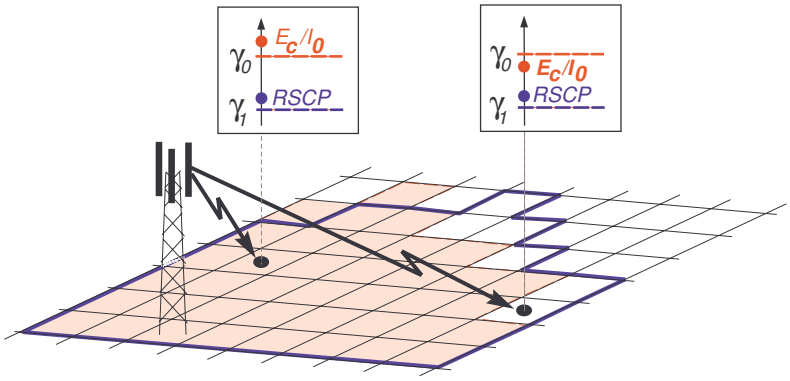


Figure 3.1: Modeling CPICH coverage.

$$\mathcal{I}_j = \{i \in \mathcal{I} : P_{ij} \leq \Pi_i^{max}\} \ . \tag{3.6}$$

For each cell i, we define a set $\mathcal{J}_i \subseteq \mathcal{J}$ that contains all bins that may be covered by the cell, i.e.,

$$\mathcal{J}_i = \{j \in \mathcal{J} : P_{ij} \leq \Pi_i^{max}\} \ . \tag{3.7}$$

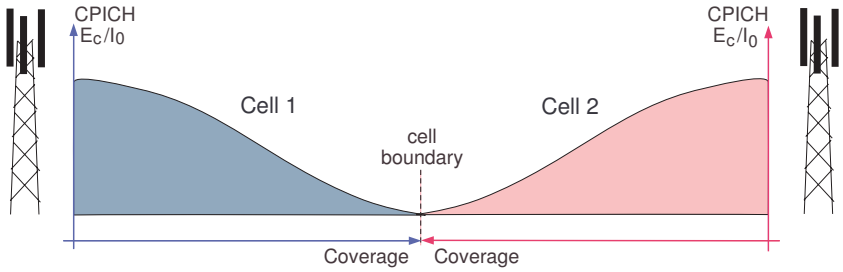
In the next section we demonstrate a simple way of handling cell overlap in the preprocessing step in order to increase the soft handover probability.

3.2 Ensuring Smooth Handover by Adjusting Power Gain Parameters

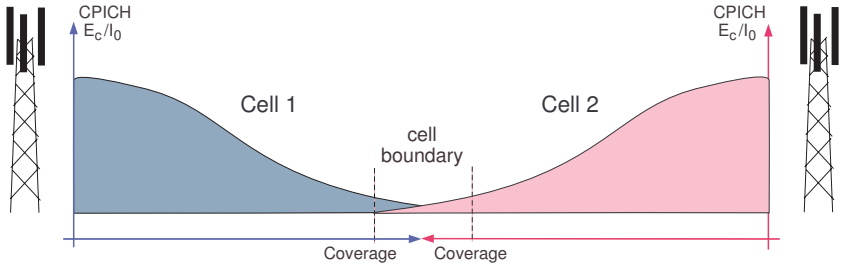
Handover is one of the essential means to support user mobility in a mobile communications network. The basic concept is simple: when the user moves from the coverage area of one cell to another, a new connection with the latter has to be set up and the connection with the old cell may be released. Soft handover is a feature specific in CDMA networks. It allows a user to have two (or even more) simultaneous connections with base stations and therefore, when moving from one cell to another, a new connection can be set up before the old one is released. To enable this procedure a certain cell overlap is necessary. In this section, we introduce the concept of *smooth handover* which implies the existence of some necessary cell overlap to make possible a smooth transition from one cell to another one by means of soft handover, i.e., without breaking the data session.

In Section 3.1, there has been presented a model for full coverage of the service area. Full coverage, however, does not necessarily ensure smooth handover. For example, consider two adjacent bins served by two different cells, for which the E_c/I_0 of each of the two CPICH signals is good in its respective bin, but very poor in the other. A mobile terminal that moves from one bin into the other, crossing the boundary of its home cell, may have difficulties in detecting the pilot signal of the other cell in time. When this occurs, there is a risk of dropping a call or interrupting a data session. A simple approach for ensuring cell overlap in the preprocessing step without modifying the model presented is presented in Section 3.1.

To facilitate smooth handover, a mobile terminal should be able to detect the pilot signal of the cell to which handover will take place, before it leaves its current home cell. For this purpose, the pilot power levels should be set such that, for the above example, the pilot



(a) Two cells that do not overlap (hard handover)



(b) Two overlapping cells (smooth handover)

Figure 3.2: The handover operation and cell overlap.

signal of a cell not only covers its own bin, but also provides some coverage in the adjacent bin served by the other cell (see Figure 3.2). One way to provide this kind of coverage is to require that the CPICH E_c/I_0 is above γ_0 in bins adjacent to the current cell. However, this may lead to an unreasonably (and unnecessarily) large amount of pilot power, and a risk of high pilot pollution.

Instead, we handle smooth handover by requiring that, if two adjacent bins belong to different cells, both pilot signals have E_c/I_0 values of at least γ_0 at the boundary of the two bins. This requirement increases the likelihood of being in soft or softer handover for mobile terminals at cell boundaries. Modeling this constraint would require prediction of signal propagation at bin boundaries. Such predictions are not available in our system model. (An implicit assumption in the model presented in Section 3.1 is that, for every bin, the power gain of a cell is identical in the entire bin.) However, it is reasonable to assume that, for two adjacent bins, the power gain at their boundary is somewhere between the gain values of the two bins. Therefore, we use the average value of the two power gain values as the power gain at the bin boundary.

Consider cell i and two adjacent bins j and j_1. If P_i^{CPICH} meets the E_c/I_0 target in both bins, or in none of the two, the aforementioned constraint of smooth handover does not apply. Assume that cell i has a sufficiently high CPICH E_c/I_0 in bin j but not in bin j_1, and that $g_{ij} > g_{ij_1}$. To enable smooth handover for mobile terminals moving from j_1 into j, i.e., to slightly extend the coverage of cell i to the near-border area in bin j, the strength of the received CPICH signal is calculated using the average value of g_{ij} and g_{ij_1}. Thus, the CPICH E_c/I_0 in j becomes as follows,

$$\frac{\frac{g_{ij}+g_{ij_1}}{2}\cdot P_i^{CPICH}}{I_j} = \frac{\frac{g_{ij}+g_{ij_1}}{2}\cdot P_i^{CPICH}}{\sum_{l\in\mathcal{I}} g_{lj}\eta_l^{DL}P_l^{max}+\nu_j+I_j^A}\,. \tag{3.8}$$

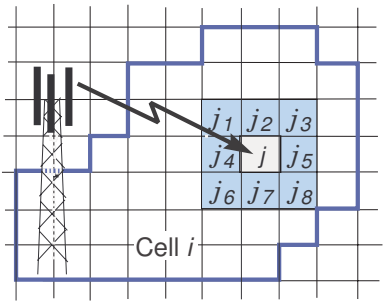


Figure 3.3: Set $\mathcal{A}_j$ of adjacent bins for bin j.

Note that, the interference computation in equation (3.8) uses the power gain of bin j, i.e., the same as in equation (3.2), not the average power gain. The reason for this is the following. The pilot power levels are planned for the scenario of worst-case interference – using the average power gain in the denominator of (3.8) would lead to less interference and thus jeopardize full coverage.

To formalize the smooth handover requirement, we use $\mathcal{A}_j$ to denote the set of adjacent bins of bin j. In Figure 3.3, the set of bins adjacent to bin j consists of eight elements, $\mathcal{A}_j = \{j_1, j_2, j_3, j_4, j_5, j_6, j_7, j_8\}$ and the cell boundary is marked with a thick solid line. Note that the set of adjacent bins $\mathcal{A}_j$ can be smaller than eight elements for j along the service area edges.

If cell i satisfies the CPICH E_c/I_0 threshold in bin j, then the new CPICH E_c/I_0 formula applies to all bins in $\mathcal{A}_j$. For convenience, we introduce notation $\bar{g}_{ij}$ to represent the new, adjusted power gain for bin j, that is,

$$\bar{g}_{ij} = \min\left\{g_{ij}, \min_{j' \in \mathcal{A}_j} \frac{g_{ij} + g_{ij'}}{2}\right\} = \min\left\{g_{ij}, \frac{g_{ij} + \min_{j' \in \mathcal{A}_j} g_{ij'}}{2}\right\} . \tag{3.9}$$

We can then formalize the *s*mooth handover constraint as follows,

$$\frac{\bar{g}_{ij} P_i^{CPICH}}{\sum_{l \in \mathcal{I}} g_{lj} \eta_l^{DL} P_l^{max} + \nu_j + I_j^A} \geq \gamma_0 . \tag{3.10}$$

Observe that (3.10) is always at least as strong as (3.1).

Ideally, we would like to ensure the smooth handover constraint only for cell boundary bins. This, however, is difficult in general and impossible in the preprocessing step since the cell boundaries are a part of the solution. Therefore, constraint (3.10) has to be applied for any bin j and therefore may be more pessimistic than necessary. This is because adjusting power gain for any of the bins, including those in the interior part of a cell, may theoretically impact the cell pilot power. However, similar to that the cell pilot power (without adjusting power gains) is typically defined by bins along the cell boundary, the optimal pilot power in a cell for adjusted power gain values is also most likely to be defined by the cell boundary bins. Therefore, it is reasonable to expect that the adjusted power gains in the interior part of the cell have no significant impact on the cell pilot power, which implies a small amount of overestimation introduced by the smooth handover constraint.

To ensure that the CPICH signal is correctly decoded at the cell border, the adjusted power gain value needs to be included in the RSCP requirement as well, i.e., the following must hold,

$$\bar{g}_{ij} P_i^{CPICH} \geq \gamma_1 . \tag{3.11}$$

Thus, with (3.8) and (3.11), the minimum CPICH needed in cell i in order to cover bin j can be found as follows,

$$P_{ij} = \max\left\{\frac{\gamma_1}{\bar{g}_{ij}}, \frac{\gamma_0}{\bar{g}_{ij}} \cdot \left(\sum_{l\in\mathcal{I}} g_{lj}\eta_l^{DL} P_l^{max} + \nu_j + I_j^A\right)\right\} , \tag{3.12}$$

where $\bar{g}_{ij}$ is defined by (3.9).

3.3 Optimization Problem

Assuming that P_i^{CPICH} are continuous, the basic pilot power optimization problem can be stated as follows.

Objective

- Find a vector $P^{CPICH} = \{P_i^{CPICH}, i \in \mathcal{I}\}$ that minimizes the total amount of pilot power in the network, i.e., $\sum_{i\in\mathcal{I}} P_i^{CPICH} \rightarrow \min$.

Constraints

- Full coverage is to be achieved, i.e., for each bin $j \in \mathcal{J}$, there must exist at least one cell i satisfying $P_i^{CPICH} \geq P_{ij}$.
- The pilot power of any cell i is within a given interval, i.e., $\Pi_i^{min} \leq P_i^{CPICH} \leq \Pi_i^{max}$.

We denote this problem by M1. The theoretical computational complexity of M1 is formalized in Theorem 3.1.

Theorem 3.1. *M1 is $\mathcal{NP}$-hard.*

Proof. We show that any instance of the minimum-cost set covering problem (which is $\mathcal{NP}$-hard) can be transformed to an instance of M1. Consider an instance of the minimum-cost set covering problem, where $\mathcal{J}$ is a ground set of elements, and $\{\mathcal{J}_i, i \in \mathcal{I}\}$ is a collection of subsets of $\mathcal{J}$. We assume that each element is contained in at least two subsets since the elements that can be covered by a single subset can be eliminated in preprocessing. Each subset $\mathcal{J}_i$ is associated with a cost c_i. A set of binary parameters $\{a_{ji}, j \in \mathcal{J}, i \in \mathcal{I}\}$ is used to denote whether subset $\mathcal{J}_i$ contains element j or not, i.e., a_{ji} is one if $j \in \mathcal{J}_i$, and zero otherwise. The objective of the minimum-cost set covering problem is to select a subset of $\mathcal{I}$ such that the corresponding subsets $\mathcal{J}_i$ have the minimum total cost and include all elements from the ground set.

The corresponding instance of M1 has a set of cells $\mathcal{I}$ and a set of bins $\mathcal{J} \cup \mathcal{J}'$, where $\mathcal{J}' = \bigcup_{i\in\mathcal{I}} \mathcal{J}_i'$ is the set of bins that can be covered by only one cell, and $\mathcal{J}_i'$ is the set of bins that can be covered by cell i only. Let $|\mathcal{J}_i'| = 1, i \in \mathcal{I}$. The minimum and the maximum CPICH power levels in cell i are defined as $\Pi_i^{min} = \epsilon$ $(0 < \epsilon < \min_{i\in\mathcal{I}} c_i)$, and $\Pi_i^{max} = c_i + \epsilon$, respectively. For every cell $i \in \mathcal{I}$ and every bin $j \in \mathcal{J} \cup \mathcal{J}_i'$, we define the minimum CPICH power in the cell needed to cover j as follows,

$$P_{ij} = \begin{cases} \epsilon & \text{if } j \in \mathcal{J}_i' , \\ c_i + \epsilon & \text{if } j \in \mathcal{J}_i , \\ \Pi_i^{max} + \epsilon & \text{if } j \notin \mathcal{J}_i \cup \mathcal{J}_i' . \end{cases} \tag{3.13}$$

The CPICH power P_i^{CPICH} of cell i is thus either $c_i + \epsilon$, if it covers the bins in set $\mathcal{J}_i$, or ϵ otherwise. This is because by definition, the CPICH power of cell i is equal to the maximum P_{ij} value among the covered bins.

The above transformation is clearly polynomial. Moreover, a feasible solution to the M1 instance is also feasible to the minimum-cost set covering instance, and vice versa. Finally, the objective function of M1 equals the objective function of the minimum-cost set covering problem plus $|\mathcal{I}| \cdot \epsilon$. Hence the conclusion. □

Theorem 3.1 suggests that polynomial-time methods that can provide an exact solution for all data sets to this problem do not exist unless a polynomial-time algorithm for at least one $\mathcal{NP}$-complete problem is found. Our goal is to design algorithms to obtain near-optimal solutions to M1 with reasonable computing effort even for large-scale networks.

3.4 Two Ad Hoc Solutions

3.4.1 Uniform Pilot Power

By the uniform pilot power approach, all cells have the same pilot power level. We use P^U to denote the pilot power used by all cells in the solution of uniform pilot power. A necessary condition for covering bin j is that P^U is at least as big as the minimum of P_{ij} among cells in $\mathcal{I}_j$. For bin j, this condition is formulated as follows,

$$P^U \geq P^j \ , \tag{3.14}$$

where P^j is the minimum pilot power required to cover bin j by at least one cell, i.e.,

$$P^j = \min_{i \in \mathcal{I}_j} P_{ij} \ . \tag{3.15}$$

To provide full coverage in the network, this condition must hold for every bin, leading to the following inequality,

$$P^U \geq \max_{j \in \mathcal{J}} \min_{i \in \mathcal{I}_j} P_{ij} \ . \tag{3.16}$$

We further observe that, if the pilot power levels are set to $\max_{j \in \mathcal{J}} \min_{i \in \mathcal{I}_j} P_{ij}$ in all cells, every bin is covered by at least one pilot signal. Since there is no reason to use more power than it is needed, it follows directly that we can change (3.16) to equality, e.g.,

$$P^U = \max_{j \in \mathcal{J}} \min_{i \in \mathcal{I}_j} P_{ij} \ . \tag{3.17}$$

The total pilot power for the solution of uniform pilot power is therefore $|\mathcal{I}| \cdot P^U$.

Using uniform pilot power is efficient in simple scenarios. However, as it will be shown later, for an in-homogeneous planning situation, applying the same uniform CPICH power approach results in unnecessarily high power consumption for pilot signals.

3.4.2 Gain-based Pilot Power

A second ad hoc solution for setting pilot power is to assign cells to bins based on power gain values. In this solution, a bin is always covered by a cell with the maximum power gain. For bin j, we use $c(j)$ to denote the cell that maximizes the power gain among all cells, i.e.,

$$c(j) = \arg \max_{i \in \mathcal{I}_j} g_{ij} \ . \tag{3.18}$$

Note that, if the total transmit power P_i^{Tot} is the same for all cells, then for any bin, the cell with the maximum power gain is also the cell with the minimum required power level (this follows from (3.3)). In this case, $c(j)$ can be equivalently defined as

$$c(j) = \arg \min_{i \in \mathcal{I}_j} P_{ij} \ . \tag{3.19}$$

Choosing the cell with the maximum gain for every bin, and setting the power levels accordingly, we obtain a solution in which all bins are covered. In this solution, the pilot power of cell i equals

$$P_i^G = \max_{\substack{j \in \mathcal{J}_i: \\ c(j)=i}} P_{ij} \ . \tag{3.20}$$

Thus, the total pilot power in the network is $\sum_{i \in \mathcal{I}} P_i^G$.

The gain-based pilot power is quite intuitive for a network planner. In fact, it significantly outperforms the solution of uniform pilot power. However, our numerical experiments show that this solution can still be quite far from the optimum, especially for large networks.

3.5 Mathematical Programming Formulations

In this section, we present several mathematical programming formulations for problem M1. First, we present a very straightforward MIP formulation which, as it will be shown later, is not efficient from the computational point of view. Then, we present two enhanced integer programming formulations that allow us to tackle the problem even for large networks.

3.5.1 Cell-bin Formulation

We use a set of continuous power variables $\{p_i, i \in \mathcal{I}\}$ and a set of binary coverage variables $\{x_{ij}, j \in \mathcal{J}, i \in \mathcal{I}_j\}$, assuming that the lower and the upper pilot power limits have been ensured in the preprocessing step. The two sets of variables are defined as follows.

$p_i =$ The pilot power of cell i,

$$x_{ij} = \begin{cases} 1 & \text{if cell } i \text{ covers bin } j, \text{ i.e., if the } E_c/I_0 \text{ and RSCP conditions hold for} \\ & \text{bin } j \text{ with respect to cell } i, \\ 0 & \text{otherwise.} \end{cases}$$

The cell-bin formulation of problem M1 (presented in Section 3.3) is presented below.

$$\text{[M1-CBF]} \quad P^* = \min \sum_{i \in \mathcal{I}} p_i \tag{3.21a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_j} x_{ij} \geq 1 \qquad j \in \mathcal{J} \tag{3.21b}$$

$$P_{ij} x_{ij} \leq p_i \qquad j \in \mathcal{J}, i \in \mathcal{I}_j \tag{3.21c}$$

$$x_{ij} \in \{0, 1\} \qquad j \in \mathcal{J}, i \in \mathcal{I}_j \tag{3.21d}$$

$$p_i \in \mathbb{R}_+ \qquad i \in \mathcal{I} \tag{3.21e}$$

In M1-CBF, constraints (3.21b) ensure full coverage. By constraints (3.21c), pilot power level of cell i is greater than or equal to P_{ij} for any bin j it covers. The optimal solution of M1-CBF is the optimal solution to M1, i.e, $P_i^{CPICH} = p_i, \forall i \in \mathcal{I}$.

M1-CBF is a straightforward linear mixed-integer formulation of problem M1. From a computational point of view, however, this formulation is not efficient. In particular, the LP relaxation of M1-CBF is very weak, i.e., the LP optimum is often far away from the integer optimum.

3.5.2 Enhanced Formulations

To avoid the aforementioned weakness of formulation M1-CBF, we enhance the formulation considering that the optimal pilot power of any cell belongs to a discrete set. This is due to the assumption by which the transmit power is set to the minimum level necessary to fulfil

the E_c/I_0 target. The observation allows us to enumerate all possible pilot power levels for each cell, which can be done through the use of binary variables.

Let $\mathcal{L}_i^{CPICH} = \{L_{ik}, k \in \mathcal{K}_i\}$ be the set of all possible CPICH power levels in cell i such that each element L_{ik} is equal to P_{ij} for some j. Set $\mathcal{K}_i$ consists of power level indices and is defined as follows, $\mathcal{K}_i = \{1, \ldots, |\mathcal{L}_i^{CPICH}|\}$. The elements of $\mathcal{L}_i^{CPICH}$ are distinct and therefore $|\mathcal{L}_i^{CPICH}| \leq |\mathcal{J}_i|$.

The set $\mathcal{L}_i^{CPICH}$ can also be constructed by the network planner by discretizing the interval $[\Pi_i^{min}, \Pi_i^{max}]$ on either linear or logarithmic scale. The problem remains $\mathcal{NP}$-hard, but the size becomes smaller. A drawback of such a problem reduction is that the obtained solution is most likely to be suboptimal with respect to the original problem.

Below are two equivalent enhanced formulations based on the observation that possible pilot power values are known since they form sets $\mathcal{L}_i^{CPICH}, i \in \mathcal{I}$. The first formulation, obtained by the *direct assignment* approach, assigns to cell i exactly one power level chosen from $\mathcal{L}_i^{CPICH}$. The second formulation uses the *incremental power* approach. The two approaches are described in more detail below.

Direct assignment approach

We define a set of binary variables $\{y_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ as follows,

$$y_{ik} = \begin{cases} 1 & \text{if pilot power level } L_{ik} \text{ is used in cell } i, \\ 0 & \text{otherwise.} \end{cases}$$

If cell i uses the k^{th} power level, then bin j is covered by i if and only if $P_{ij} \leq L_{ik}$. This information is represented by the following set of indication parameters.

$$a_{ijk} = \begin{cases} 1 & \text{if bin } j \text{ is covered by the pilot signal of cell } i, \text{ provided that} \\ & \text{the pilot power of cell } i \text{ equals } L_{ik}, \\ 0 & \text{otherwise.} \end{cases}$$

An enhanced formulation of M1 is presented below.

$$\text{[M1-EFDA]} \quad P^* = \min \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} L_{ik} y_{ik} \tag{3.22a}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}_j} \sum_{k \in \mathcal{K}_i} a_{ijk} y_{ik} \geq 1 \qquad j \in \mathcal{J} \tag{3.22b}$$

$$\sum_{k \in \mathcal{K}_i} y_{ik} \leq 1 \qquad i \in \mathcal{I} \tag{3.22c}$$

$$y_{ik} \in \{0, 1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{3.22d}$$

In M1-EFDA, the coverage requirement is modeled by (3.22b), and constraints (3.22c) state that at most one power level is selected for each cell.

Given a solution to M1-EFDA, the pilot power setting in the network is

$$P_i^{CPICH} = \sum_{k \in \mathcal{K}_i} L_{ik}\ y_{ik}, \qquad i \in \mathcal{I}\ . \tag{3.23}$$

Incremental power approach

In the incremental power approach we utilize the fact that if a set of bins is covered by a cell with some pilot power level, then the set is also covered if a higher CPICH power level is used in the cell. Assume that set $\mathcal{K}_i$ contains indices of sorted in ascending order power

levels from set $\mathcal{L}_i^{CPICH}$. Now consider the sorted sequence of power levels in cell i in the following incremental fashion:

$$\begin{aligned} L_{i1}^I &= L_{i1} \ , \\ L_{i2}^I &= L_{i2} - L_{i1} \ , \\ &\dots \\ L_{ik}^I &= L_{ik} - L_{i(k-1)} \ , \\ &\dots \\ L_{i|\mathcal{K}_i|}^I &= L_{i|\mathcal{K}_i|} - L_{i(|\mathcal{K}_i|-1)} \ . \end{aligned}$$

The pilot power of cell i can thus be expressed as $\sum_{k\in\mathcal{K}_i} L_{ik}^I x_{ik}$, and the total amount of pilot power is $\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L_{ik}^I x_{ik}$, where x_{ik} is a binary variable from set $\{x_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$, defined as

$$x_{ik} = \begin{cases} 1 & \text{if cell } i \text{ uses pilot power of at least } L_{ik}, \\ 0 & \text{otherwise.} \end{cases}$$

For each bin j we define $\kappa(i,j)$ as the index of the minimum power level of cell i with which the bin may be covered by the cell, i.e.,

$$\kappa(i,j) = \arg \min_{\substack{k\in\mathcal{K}_i: \\ L_{ik}\geq P_{ij}}} L_{ik} \ . \tag{3.24}$$

With the above notation, the enhanced formulation using incremental power can be presented as follows.

$$\text{[M1-EFIP]} \quad P^* = \min \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L_{ik}^I x_{ik} \tag{3.25a}$$

$$\text{s.t.} \sum_{i\in\mathcal{I}_j} x_{i\kappa(i,j)} \geq 1 \qquad j \in \mathcal{J} \tag{3.25b}$$

$$x_{i(k-1)} \geq x_{ik} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \setminus \{1\} \tag{3.25c}$$

$$x_{ik} \in \{0,1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{3.25d}$$

In M1-EFIP, constraints (3.25b) model the full coverage requirement and constraints (3.25c) model sequences of non-decreasing CPICH power levels.

For a solution in x, the pilot power setting in the network is as follows,

$$P_i^{CPICH} = \sum_{k\in\mathcal{K}_i} L_{ik}^I x_{ik}, \qquad i \in \mathcal{I} \ . \tag{3.26}$$

Note that in an optimal solution of either of the enhanced formulations, the CPICH power of a cell is at least at the required minimum Π_i^{min} only if it covers at least one bin (provided that P_{ij}-values have been adjusted accordingly in the preprocessing step). Otherwise it is zero. Ensuring the minimum CPICH power in all cells in the optimal solutions is trivial and can be applied to either of the enhanced formulations: one can subtract Π_i^{min} from every P_{ij}-value (provided that $P_{ij} \geq \Pi_i^{min}, \forall i \in \mathcal{I}, j \in \mathcal{J}_i$) and then add a constant value of $\sum_{i\in\mathcal{I}} \Pi_i^{min}$ to the optimal value of the objective function. In our numerical experiments, we also apply this adjustment such that CPICH power in all cells is within $[\Pi_i^{min}, \Pi_i^{max}]$.

Observe that formulations M1-EFDA and M1-EFIP can be easily transformed to each other. Furthermore, their LP relaxations are equivalent in terms of provided LP bounds, which is formally stated in the following theorem.

Theorem 3.2. *The lower bounds provided by the LP relaxation of M1-EFDA and M1-EFIP are equal.*

Proof. Let us denote the LP relaxation of M1-EFIP and M1-EFDA by M1-EFIP-LP and M1-EFDA-LP, respectively. Consider an optimal solution $\bar{\mathrm{x}} = \{\bar{x}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ to M1-EFIP-LP. Without loss of generality, we assume that index $k \in \mathcal{K}_i$ denotes the largest index of the non-zero x-variable for cell i. This index corresponds to a certain power level L_{ik} in the ordered (in ascending order) set $\mathcal{L}_i^{CPICH}$. Let $k = \kappa(i, j_i^*)$ for some bin j_i^*. Note that by definition, $\bar{x}_{ik} = 0$ for all $k > \kappa(i, j_i^*)$.

Let $\bar{\mathrm{y}} = \{\bar{y}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ be a solution to M1-EFDA-LP defined as follows,

$$\bar{y}_{ik} = \begin{cases} \bar{x}_{ik} - \bar{x}_{i(k+1)} & \text{if } k < \kappa(i, j_i^*) \\ \bar{x}_{ik} & \text{if } k = \kappa(i, j_i^*) \\ 0 & \text{if } k > \kappa(i, j_i^*) \end{cases} . \tag{3.27}$$

The solution above is feasible to M1-EFDA-LP since the condition $\bar{y}_{ik} \in [0, 1], i \in \mathcal{I}, k \in \mathcal{K}_i$ is satisfied by (3.27), and constraints (3.22c) are satisfied because the following holds

$$\sum_{k \in \mathcal{K}_i} \bar{y}_{ik} = \sum_{\substack{k \in \mathcal{K}_i: \\ k \le \kappa(i, j_i^*)}} \bar{y}_{ik} = \bar{x}_{i1} \le 1 , \tag{3.28}$$

and constraints (3.22b) are satisfied because of the following valid equations

$$\sum_{i \in \mathcal{I}_j} \sum_{k \in \mathcal{K}_i} a_{ijk} \bar{y}_{ik} = \sum_{i \in \mathcal{I}_j} \sum_{\substack{k \in \mathcal{K}_i: \\ k \le \kappa(i, j_i^*)}} a_{ijk} \bar{y}_{ik} = \sum_{i \in \mathcal{I}_j} \sum_{\substack{k \in \mathcal{K}_i: \\ \kappa(i,j) \le k \le \kappa(i, j_i^*)}} \bar{y}_{ik} = \sum_{\substack{i \in \mathcal{I}_j: \\ \kappa(i,j) \le \kappa(i, j_i^*)}} \bar{x}_{i\kappa(i,j)} \ge 1 . \tag{3.29}$$

The last inequality in (3.29) is due to that $\bar{\mathrm{x}}$ is optimal (and feasible) to M1-EFIP-LP.

Below we show that the objective function values of the two LP relaxations are equal.

$$\begin{aligned} &\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} L_{ik}^I \bar{x}_{ik} = \sum_{i \in \mathcal{I}} \sum_{\substack{k \in \mathcal{K}_i: \\ k \le \kappa(i, j_i^*)}} L_{ik}^I \bar{x}_{ik} = \sum_{i \in \mathcal{I}} \sum_{\substack{k \in \mathcal{K}_i: \\ k \le \kappa(i, j_i^*)}} (L_{ik} - L_{i(k-1)}) \bar{x}_{ik} = \\ &\sum_{i \in \mathcal{I}} \sum_{\substack{k \in \mathcal{K}_i: \\ k < \kappa(i, j_i^*)}} L_{ik} (\bar{x}_{ik} - \bar{x}_{i(k+1)}) + L_{i\kappa(i,j_i^*)} \bar{x}_{i\kappa(i,j_i^*)} = \sum_{i \in \mathcal{I}} \sum_{\substack{k \in \mathcal{K}_i: \\ k \le \kappa(i, j_i^*)}} L_{ik} \bar{y}_{ik} = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} L_{ik} \bar{y}_{ik} \end{aligned} \tag{3.30}$$

With the transformations above, we have proved that if solution $\bar{\mathrm{x}}$ is optimal to M1-EFIP-LP, then there exists some solution $\bar{\mathrm{y}}$ feasible to M1-EFDA-LP, and the objective function values of the two LP problems are equal. In other words, the optimal objective function value of M1-EFDA-LP does not exceed the optimal objective function value of M1-EFIP-LP. Next we show that the latter does not exceed the former.

Without loss of generality, we assume that $\mathcal{K}_i$ is the ordered set of power level indexes in cell i, i.e., a smaller k corresponds to a smaller amount of pilot power. Let $\tilde{\mathrm{y}} = \{\tilde{y}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ be an optimal solution to M1-EFDA-LP.

We derive a solution $\tilde{\mathrm{x}} = \{\tilde{x}_{ik} = \sum_{k' \in \mathcal{K}_i : k' \ge k} \tilde{y}_{ik'},\ i \in \mathcal{I}, k \in \mathcal{K}_i\}$ feasible to M1-EFIP-LP. By construction, solution $\tilde{\mathrm{x}}$ satisfies (3.25c) and $\tilde{x}_{ik} \in [0, 1], i \in \mathcal{I}, k \in \mathcal{K}_i$. Furthermore, the following equalities show that for any bin j, the solution also satisfies the corresponding constraint from (3.25b),

$$\sum_{i \in \mathcal{I}_j} \tilde{x}_{i\kappa(i,j)} = \sum_{i \in \mathcal{I}_j} \sum_{\substack{k' \in \mathcal{K}_i: \\ k' \ge \kappa(i,j)}} \tilde{y}_{ik'} = \sum_{i \in \mathcal{I}_j} \sum_{k' \in \mathcal{K}_i} a_{ijk'} \tilde{y}_{ik'} \ge 1 . \tag{3.31}$$

Moreover, the objective function value of M1-EFIP-LP for solution $\tilde{\mathrm{x}}$ equals the objective

function value of M1-EFDA-LP for solution $\tilde{\mathrm{y}}$:

$$\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L^I_{ik}\tilde{x}_{ik} = \sum_{i\in\mathcal{I}}\left(L_{i1}\tilde{x}_{i1} + \sum_{\substack{k\in\mathcal{K}_i:\\ k\geq 2}}(L_{ik}-L_{i(k-1)})\tilde{x}_{ik}\right) =$$

$$\sum_{i\in\mathcal{I}}\left(\sum_{k\in\mathcal{K}_i\setminus\{|\mathcal{K}_i|\}} L_{ik}(\tilde{x}_{ik}-\tilde{x}_{i(k+1)}) + L_{i|\mathcal{K}_i|}\tilde{x}_{i|\mathcal{K}_i|}\right) = \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L_{ik}\tilde{y}_{ik}\ . \qquad (3.32)$$

Thus, we have shown that the optimal objective function value of M1-EFIP-LP does not exceed that of M1-EFDA-LP. With the result of (3.30), this implies that the two optimal objective function values are equal. Hence the conclusion. □

The LP relaxation of each of the enhanced formulations presented in this section provides a sharper bound to P^* than that of the cell-bin formulation M1-CBF. The result is formalized in Theorem 3.3. Note that the LP relaxations of M1-EFDA and M1-EFIP provide true LP bounds to M1 (and also to M1-CBF) only when $P_{ij} = L_{i\kappa(i,j)}$ for all $j \in \mathcal{J}, i \in \mathcal{I}_j$, i.e., when discretization of the CPICH power range is not used in any of the formulations or the P_{ij}-values are adjusted accordingly. Otherwise, if $P_{ij} < L_{i\kappa(i,j)}$ for some i and j, the optimal solutions to the LP relaxations of M1-EFDA and M1-EFIP still give larger objective function values than that in the optimal solution to the LP relaxation of M1-CBF, but the LP solutions to M1-EFDA and M1-EFIP do not give a lower bound to M1-CBF because they solve the problem for a different instance of M1 than that considered by M1-CBF.

Theorem 3.3. *The lower bounds provided by the LP relaxations of M1-EFDA and M1-EFIP are at least as strong as that of M1-CBF. In addition, there are instances for which the lower bounds obtained from the LP relaxations of M1-EFDA and M1-EFIP are strictly better.*

Proof. Let M1-CBF-LP denote the LP relaxation of M1-CBF, M1-EFDA-LP denote the LP relaxation of M1-EFDA, and M1-EFIP-LP denote the LP relaxation of M1-EFIP. We prove the first part of the theorem for M1-EFDA as follows. Given a feasible solution to M1-EFDA-LP, we observe that

1. this solution is also feasible in M1-CBF-LP, and
2. the objective function value of M1-EFDA-LP for this solution is greater than or equal to that of M1-CBF-LP.

Consider a feasible solution, denoted by $\bar{\mathrm{y}} = \{\bar{y}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$, to M1-EFDA-LP. Consider a solution $\hat{\mathrm{x}} = \{\hat{x}_{ij}, i \in \mathcal{I}, j \in \mathcal{J}_i\}$ to M1-CBF-LP such that $\hat{x}_{ij} = \sum_{k\in\mathcal{K}_i} a_{ijk}\bar{y}_{ik}$. It is easy to verify that $\hat{\mathrm{x}}$ satisfies (3.21b). Next, we show that the total power of solution $\hat{\mathrm{x}}$ is at most as that of solution $\bar{\mathrm{y}}$.

For solution $\hat{\mathrm{x}}$, the optimal value of p_i in M1-CBF-LP is obviously $\hat{p}_i = \max_{j\in\mathcal{J}_i} P_{ij}\hat{x}_{ij}$. Assume that the maximum occurs for bin j^*, i.e., $\hat{p}_i = P_{ij^*}\hat{x}_{ij^*}$. In M1-EFDA-LP, the pilot power of cell i reads $\sum_{k\in\mathcal{K}_i} L_{ik}\bar{y}_{ik}$. The following sequence of inequalities and equalities is valid,

$$\sum_{k\in\mathcal{K}_i} L_{ik}\bar{y}_{ik} \geq \sum_{\substack{k\in\mathcal{K}_i:\\ P_{ij^*}\leq L_{ik}}} L_{ik}\bar{y}_{ik} \geq P_{ij^*}\sum_{\substack{k\in\mathcal{K}_i:\\ P_{ij^*}\leq L_{ik}}} \bar{y}_{ik} = P_{ij^*}\sum_{k\in\mathcal{K}_i} a_{ij^*k}\ \bar{y}_{ik} = P_{ij^*}\hat{x}_{ij^*} = \hat{p}_i\ . \qquad (3.33)$$

Because this holds for any cell, we have proved the first part of the theorem for M1-EFDA. By Theorem 3.2, the first part of the theorem also holds for M1-EFIP.

To show the second part of the theorem, it is sufficient to give an example. Consider two cells and four bins, where $P_{11} = 1.2, P_{12} = 0.8, P_{13} = 0.6, P_{21} = 0.6, P_{22} = 0.8$, and

$P_{24} = 0.3$. Assume also that P_{14} and P_{23} exceed the power limit (and are thus irrelevant to the discussion). In the integer optimum of M1-CBF, $x_{12} = x_{13} = x_{21} = x_{24} = 1$ or $x_{13} = x_{21} = x_{22} = x_{24} = 1$, and the total pilot power equals 1.4. M1-CBF-LP gives $\hat{x}_{11} = 0.5, \hat{x}_{12} = 0.75, \hat{x}_{13} = 1, \hat{x}_{21} = 0.6, \hat{x}_{22} = 0.25, \hat{x}_{23} = 1$, and the objective function value equals 0.9. (The relative gap is therefore 36 %.) M1-EFDA-LP and M1-EFIP-LP, on the other hand, yield the integer optimum. □

In Sections 3.6 and 3.7, two different solution approaches to M1 are presented. The first approach is based on Lagrangian relaxation, and the second solution approach is based on column generation. Both approaches can be applied to either of the presented enhanced formulations. We choose the enhanced formulation based on direct assignment (M1-EFDA) to demonstrate the column generation approach and the enhanced formulation based on incremental pilot power (M1-EFIP) to present the Lagrangian heuristic.

3.6 A Solution Approach Based on Lagrangian Relaxation

3.6.1 Algorithm Overview

The idea of Lagrangian relaxation is to relax hard constraints by bringing them into the objective function with associated Lagrange multipliers. The multipliers are not restricted in sign, if the relaxed constraints are equalities. Otherwise, they are restricted to be either nonnegative or nonpositive, depending on the inequality sign and whether this is a maximization or minimization problem. When solving a minimization problem, the Lagrangian function gives a lower bound on the optimal objective function value of the original problem (*Lagrangian bounding principle*). Therefore, to find the sharpest possible lower bound to a minimization problem, we need to solve an optimization problem where Lagrangian function is to be maximized with respect to Lagrange multipliers. The problem is referred to as *Lagrangian dual problem*.

Our approach consists of two parts. In the first part, we solve the Lagrangian multiplier problem by a subgradient optimization technique (see, for example, [8]) where in each iteration the Lagrangian relaxation is to be solved. Utilizing the problem structure, we find the optimal solution to the Lagrangian relaxation by solving a set of smaller independent subproblems (one subproblem for each cell). The first part of the algorithm provides us with a lower bound to the pilot power optimization problem. The second part of the algorithm was designed to obtain a reasonably good feasible solution and an upper bound to the original problem. This is done by applying a power-adjustment heuristic which modifies the solutions to the relaxed problem. The power-adjustment heuristic is applied in every iteration of the subgradient method. We use three stopping criteria: maximum number of subgradient steps (500 steps), dual gap less than 1 %, and maximum number of consecutive steps during which the lower bound has not been improved (50 steps).

The Lagrangian heuristic is outlined in Algorithm I.1. The algorithm requires the following input data,

- → a set of cells $\mathcal{I}$,
- → a set of bins $\mathcal{J}$,
- → a set of power level indices of each cell, $\mathcal{K}_i, i \in \mathcal{I}$,
- → incremental power parameters $\{L^I_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$,
- → subsets $\{\mathcal{J}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ of bins for which power level L_{ik} is the minimum for being covered by cell i (see equation (3.34) for a formal definition of $\mathcal{J}_{ik}$),
- → an initial feasible solution $\mathrm{x}^0 = \{x^0_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$,

Algorithm I.1 Lagrangian heuristic

Input: $\mathcal{I}, \mathcal{J}, \{\mathcal{K}_i, i \in \mathcal{I}\}, \{L^I_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}, \{J_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}, \mathrm{x}^0, N_1, N_2, gap$
Output: LB, P^*, x^*

$\mathrm{x} \Leftarrow \mathrm{x}^0$
$P \Leftarrow$ **objEFIP**(x) // Compute the objective function value
$\mathrm{x}^* \Leftarrow \mathrm{x}$
$P^* \Leftarrow P$
$\lambda \Leftarrow \{0\}^{|\mathcal{J}|}$
$LB \Leftarrow 0$
$step \Leftarrow 0$
$badStep \Leftarrow 0$
$\mu \Leftarrow 2.0$
repeat
    $\mathrm{x} \Leftarrow$ **solveLagrRel**(λ) // Solve Lagrangian subproblem
    $lb \Leftarrow LF(\mathrm{x}, \lambda)$ // Evaluate Lagrangian function
    if $lb > LB$ **then**
        $LB \Leftarrow lb$
        $badStep \Leftarrow 0$
    else
        $badStep \Leftarrow badStep + 1$
        if muDecr($badStep$) **then**
            $\mu \Leftarrow \mu/2$
        end if
    end if
    $\mathrm{x}' \Leftarrow$ **findFeas**(x) // Run the power adjustment heuristic
    $P \Leftarrow$ **objEFIP**(x') // Compute the objective function value
    if $P < P^*$ **then**
        $P^* \Leftarrow P$
        $\mathrm{x}^* \Leftarrow \mathrm{x}'$
    end if
    $E \Leftarrow 0$
    for $\forall j \in \mathcal{J}$ **do**
        $e_j \Leftarrow 1 - \sum_{\mathcal{I}_j} x_{i\kappa(i,j)}$
        $E \Leftarrow E + e_j^2$
    end for
    $\theta \Leftarrow \mu \cdot (P^* - lb)/E$
    for $\forall j \in \mathcal{J}$ **do**
        $\lambda_j \Leftarrow \max(0, \lambda_j + \theta \cdot e_j)$
    end for
    $step \Leftarrow step + 1$
until $(step == N_1) \vee (badStep == N_2) \vee ((P^* - LB)/LB \leq gap)$

→ the maximum number of iterations N_1,

→ the maximum number of consecutive and non-improving iterations N_2,

→ the target duality *gap*.

The formal definition of $\mathcal{J}_{ik}$ is as follows,

$$\mathcal{J}_{ik} = \{j \in \mathcal{J}_i : \kappa(i,j) = k\} \ . \tag{3.34}$$

The initial solution can be obtained, for example, by the gain-based approach suggested in Section 3.4.2. The output is the best found lower bound LB, best feasible solution power P^*, and the corresponding solution $\mathrm{x}^* = \{x^*_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$.

In line (11) of the algorithm, function **solveLagrRel**(λ) solves $|\mathcal{I}|$ cell-level subproblems of the Lagrangian relaxation and returns the optimal solution to the entire Lagrangian subproblem for a given vector of Lagrange multipliers λ. The solution approach is presented in more detail in Section 3.6.2. Function **findFeas**(x) is the power-adjustment heuristic. The function takes the current solution to the Lagrangian subproblem and returns a feasible solution to the original problem. The details of the power-adjustment heuristic are presented in Section 3.6.3. Initially, Lagrange multipliers are a vector of zeros of size $|\mathcal{J}|$ (see line (5)).

Function **objEFIP**(x) computes the objective function value of M1-EFIP for a given solution x. Lagrange multipliers are updated in lines (28)-(36), where we first compute a step size θ which specifies how far we move in a subgradient direction and then update the Lagrange multipliers. The step size is computed as follows [8],

$$\theta = \frac{\mu \cdot (P^* - LF(\mathrm{x}, \lambda))}{\sum_{j \in \mathcal{J}} e_j^2} \ , \tag{3.35}$$

where μ is a scalar (strictly) between zero and two, P^* is the best feasible solution found so far, $LF(\mathrm{x}, \lambda)$ is the Lagrangian function value for x and λ, and $e_j = 1 - \sum_{i \in \mathcal{I}_j} x_{i\kappa(i,j)}$ is the slack of the corresponding relaxed constraint. The denominator of (3.35) is the squared Euclidean norm of the slack vector of the relaxed constraints. In Algorithm I.1, it is denoted by E. The new value of a Lagrange multiplier λ_j is computed by

$$\lambda_j = \max\left(0, \lambda_j + \theta \cdot e_j\right) \ , \tag{3.36}$$

where e_j is the j^{th} element of the subgradient vector e. Note that the Lagrange multipliers are constrained to be nonnegative since the relaxed constraints have the sign "$\geq$".

The initial value of μ is two. It is reduced by a factor of two if the best lower bound does not increase in a specified number of consecutive iterations. The condition is checked in **muDecr**(*badStep*). In our computational experiments, the number is set to ten.

3.6.2 Lagrangian Relaxation

A Lagrangian relaxation that exploits the structure of the problem is the core of the algorithm presented in Section 3.6.1.

Consider the enhanced formulation M1-EFIP. We relax the coverage constraints (3.25b) using non-negative Lagrange multipliers, $\{\lambda_j, j \in \mathcal{J}\}$, and construct the following Lagrangian function.

$$\begin{aligned} LF(x, \lambda) = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} L^I_{ik} x_{ik} + \sum_{j \in \mathcal{J}} \left[\lambda_j \left(1 - \sum_{i \in \mathcal{I}_j} x_{i\kappa(i,j)} \right) \right] = \\ = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} L^I_{ik} x_{ik} + \sum_{j \in \mathcal{J}} \lambda_j - \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}_j} \lambda_j x_{i\kappa(i,j)} \ . \end{aligned} \tag{3.37}$$

By the definitions of $\mathcal{I}_j$ and $\mathcal{J}_i$ (see Section 3.1) the following equality holds,

$$\sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}_j} \lambda_j x_{i\kappa(i,j)} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} \lambda_j x_{i\kappa(i,j)} \ . \tag{3.38}$$

Moreover, with $\mathcal{J}_{ik}$ given, we get

$$\sum_{j \in \mathcal{J}_i} \lambda_j x_{i\kappa(i,j)} = \sum_{k \in \mathcal{K}_i} \sum_{j \in \mathcal{J}_{ik}} \lambda_j x_{ik} \ . \tag{3.39}$$

As a result, with (3.38) and (3.39), the Lagrangian function (3.37) reads

$$LF(x,\lambda) = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}_i} \left(L^I_{ik} - \sum_{j\in\mathcal{J}_{ik}} \lambda_j \right) x_{ik} + \sum_{j\in\mathcal{J}} \lambda_j \ . \tag{3.40}$$

Thus, the relaxed problem is to minimize $LF(x,\lambda)$ in x over the set defined by the remaining (not relaxed) constraints of M1-EFIP, i.e.,

$$\begin{aligned} \text{[M1-EFIP-R1]} \quad LF(x,\lambda) \quad &\to \min \\ \text{s.t.} \quad x_{i(k-1)} &\geq x_{ik} && i\in\mathcal{I}, k\in\mathcal{K}_i\setminus\{1\} \\ x_{ik} &\in \{0,1\} && i\in\mathcal{I}, k\in\mathcal{K}_i \end{aligned}$$

We decompose problem M1-EFIP-R1 into $|\mathcal{I}|$ independent subproblems (one per cell) and solve each subproblem with respect to the x-variables. We use M1-EFIP-R1$_i$ to denote the subproblem to be solved for cell i. The subproblems are presented below.

$$\begin{aligned} \text{[M1-EFIP-R1}_i\text{]} \quad \sum_{k\in\mathcal{K}_i} \left(L^I_{ik} - \sum_{j\in\mathcal{J}_{ik}} \lambda_j \right) x_{ik} \quad &\to \min && \text{(3.42a)} \\ \text{s.t.} \quad x_{i(k-1)} &\geq x_{ik} && k\in\mathcal{K}_i\setminus\{1\} \\ x_{ik} &\in \{0,1\} && k\in\mathcal{K}_i \end{aligned}$$

A simple way to solve the i^{th} subproblem M1-EFIP-R1$_i$ is to find, for cell i,

$$k^* = \arg\min_{q\in\mathcal{K}_i} \left\{ \sum_{k\in\mathcal{K}_i : k\leq q} \left(L^I_{ik} - \sum_{j\in\mathcal{J}_{ik}} \lambda_j \right) \right\}, \tag{3.43}$$

and then assign x_{ik} as follows,

$$x_{ik} = \begin{cases} 1 & \text{if } k\in[1,k^*], \\ 0 & \text{if } k\in[k^*+1, |\mathcal{K}_i|\,]. \end{cases} \tag{3.44}$$

Solving the $|\mathcal{I}|$ subproblems gives us a solution which, however, does not necessarily satisfy constraints (3.25b). To find a feasible solution, we apply the primal heuristic procedure discussed in the next section.

3.6.3 A Primal Heuristic Procedure

Our primal heuristic procedure is an implementation of function **findFeas**(x) in Algorithm I.1, and it consists of the following two phases:

- Increase the coverage area of cells until the coverage constraint is satisfied;
- Reduce the cell overlap maintaining the same coverage degree.

The first phase aims to adjust the obtained solution to the relaxed problem M1-EFIP-R1 to a feasible one, whereas the goal of the second phase of the heuristic procedure is to improve the solution by reducing cell overlap in the network. The two phases are performed sequentially. However, to improve the result, the second phase may be applied twice, i.e., before and after the first phase. Below we present two algorithms that can be used in the first and second phase, respectively.

Algorithm I.2 Coverage adjustment in the Lagrangian heuristic

Input: $\mathcal{J}, \mathcal{I}, \{\mathcal{I}_j, j \in \mathcal{J}\}, \{\mathcal{K}_i, i \in \mathcal{I}\}, \{L_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$, x

Output: x

1: $\bar{\mathcal{J}} \Leftarrow$ **sortA**$(\{|\mathcal{I}_j|, j \in \mathcal{J} : \sum_{i \in \mathcal{I}_j} x_{i\kappa(i,j)} < 1\})$ // Sort uncovered bins in ascending order by the number of possible covering cells
2: **for** $\forall i \in \mathcal{I}$ **do**
3: $\quad \bar{q}_i \Leftarrow \sum_{k \in \mathcal{K}_i} x_{ik}$ // Find the power level index (in the ordered set) used in cell i
4: **end for**
5: **while** $|\bar{\mathcal{J}}| > 0$ **do**
6: $\quad j \Leftarrow$ **head**$(\bar{\mathcal{J}})$ // Take the first element
7: $\quad i \Leftarrow \arg\min_{l \in \mathcal{I}_j} \left(L_{l\kappa(l,j)} - L_{l\bar{q}_i}\right)$ // Find the covering cell with minimum additional power
8: $\quad \mathcal{J}' \Leftarrow \{j' \in \bar{\mathcal{J}} : \bar{q}_i < \kappa(i,j') \leq \kappa(i,j)\}$ // Find bins covered in this iteration
9: $\quad \bar{\mathcal{J}} \Leftarrow \bar{\mathcal{J}} \setminus \mathcal{J}'$
10: $\quad$ **for** $\forall j' \in \mathcal{J}'$ **do**
11: $\quad\quad x_{ij'} \Leftarrow 1$
12: $\quad$ **end for**
13: $\quad \bar{q}_i \Leftarrow \kappa(i,j)$
14: **end while**

An Algorithm for Adjusting the Lagrangian Subproblem Solution to Feasibility

The algorithm uses as input the optimal solution to the relaxed problem M1-EFIP-R1 and adjusts it to a feasible solution. The main idea of the algorithm is to check the coverage state of every bin and, if it is uncovered, to increase the amount of pilot power of some cell to provide coverage.

If there are several cells that may cover an uncovered bin, we choose the cell with the least additional power[3] needed to cover this bin. Note that covering an uncovered bin j may result in covering some other bins also, i.e., those for which the pilot power is less than or equal to what is needed to cover j. Thus, the order in which the uncovered bins are chosen may affect the final result significantly. Obviously, the bins that may be covered by only one or two cells have a lower probability of being covered when the pilot power of some cell is increased. To allow the coverage of such bins first, we use a sequence where the uncovered bins are ordered in ascending order in the number of cells which may provide coverage.

The algorithm flow is presented in Algorithm I.2. The input of the algorithm is listed below,

- → a set of bins $\mathcal{J}$,
- → a set of cells $\mathcal{I}$,
- → a set of possible covering cells $\{\mathcal{I}_j, j \in \mathcal{J}\}$,
- → a set of power level indices for each cell, $\{\mathcal{K}_i, i \in \mathcal{I}\}$,
- → an ordered (in ascending order) set of power levels $\{L_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$,
- → a solution x that needs coverage adjustment.

[3] *Additional power* for cell i and bin j is the amount of power by which the pilot power in cell i must increase in order to cover bin j.

Algorithm I.3 Overlap reduction in the Lagrangian heuristic

Input: $\mathcal{I}, \{\mathcal{K}_i, i \in \mathcal{I}\}, \{L^I_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}, \{\mathcal{J}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$, x
Output: x

$\bar{\mathcal{I}} \Leftarrow$ **sortD**$(\{\sum_{k\in\mathcal{K}_i} L^I_{ik} x_{ik}, i \in \mathcal{I}\})$ // Sort cells in descending order by their power
for $\forall i \in \mathcal{I}$ **do**
    $\bar{q}_i \Leftarrow \sum_{k\in\mathcal{K}_i} x_{ik}$ // Find the power level index (in the ordered set) used in cell i
end for
while $|\bar{\mathcal{I}}| > 0$ **do**
    $i \Leftarrow$ **head**$(\bar{\mathcal{I}})$ // Take the first element
    repeat
        for $\forall j \in \mathcal{J}_{i\bar{q}_i}$ **do**
            covered $\Leftarrow (\sum_{l\in\mathcal{I}_j\setminus\{i\}} x_{l\kappa(l,j)} \geq 1)$ // Check if j is covered by other cells
            if $\neg$*covered* **then**
                break
            end if
        end for
        if *covered* **then**
            $x_{i\bar{q}_i} \Leftarrow 0$
            $\bar{q}_i \Leftarrow \bar{q}_i - 1$
        end if
    until $\neg$*covered*
    $\bar{\mathcal{I}} \Leftarrow \bar{\mathcal{I}} \setminus \{i\}$
end while

An Algorithm for Reducing Cell Overlap

The main idea of the second phase of the power adjustment heuristic is to reduce the size of unnecessary cell overlap areas in the network. Our approach is to examine each cell and reduce its pilot power as much as possible without coverage degradation. The algorithm (see Algorithm I.3) goes through all cells one by one. To prioritize pilot power reduction in cells with high CPICH power level, we sort all cells by their pilot power levels in descending order. The algorithm requires the following input data,

- → a set of cells $\mathcal{I}$,
- → a set of power level indices $\{\mathcal{K}_i, i \in \mathcal{I}\}$,
- → a set of incremental power levels $\{L^I_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$,
- → subsets of bins $\{\mathcal{J}_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ defined by (3.34),
- → a solution x that needs cell overlap reduction.

3.7 A Solution Approach Based on Column Generation

In this section, the second solution approach to the pilot power optimization problem M1 is presented. The algorithm is based on column generation embedded into an iterative rounding procedure. To demonstrate the solution algorithm, we chose formulation M1-EFDA, although it be applied to both enhanced formulations discussed in Section 3.5.2.

3.7.1 The Column Generation Method

Let's consider the LP-relaxation of M1-EFDA, where the integrality constraints (3.22d) are relaxed and replaced by the following constraints:

$$y_{ik} \geq 0, \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{3.45}$$

The problem with objective function (3.22a) subject to constraints (3.22b), (3.22c), and (3.45), is denoted by M1-EFDA-LP. In a column generation context, this is also referred to as the *master problem*.

In column generation, a subset $\mathcal{K}'_i \subseteq \mathcal{K}_i$ is used instead of $\mathcal{K}_i$. Usually, the size of $\mathcal{K}'_i$ is much smaller than that of $\mathcal{K}_i$. By restricting $\mathcal{K}_i$ to $\mathcal{K}'_i$ in problem M1-EFDA-LP, the following problem is obtained.

$$\text{[M1-EFDA-MAS]} \quad \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}'_i} L_{ik} y_{ik} \quad \rightarrow \min \tag{3.46a}$$

$$\text{s. t.} \sum_{i\in\mathcal{I}_j} \sum_{k\in\mathcal{K}'_i} a_{ijk} y_{ik} \geq 1 \qquad j \in \mathcal{J} \tag{3.46b}$$

$$\sum_{k\in\mathcal{K}'_i} y_{ik} \leq 1 \qquad i \in \mathcal{I} \tag{3.46c}$$

$$y_{ik} \geq 0 \qquad i \in \mathcal{I}, k \in \mathcal{K}'_i \tag{3.46d}$$

M1-EFDA-MAS is a restriction of M1-EFDA-LP, and is therefore referred to as the *restricted master problem*. We assume that M1-EFDA-MAS is feasible. (As will be clear later on, the feasibility of M1-EFDA-MAS can be easily ensured.) The optimal solution to M1-EFDA-MAS is feasible to M1-EFDA-LP. To examine whether this solution is also optimal to M1-EDFA-LP, we need to identify whether there exists a cell i and a pilot power index $k \in \mathcal{K}_i$, for which the reduced cost of the corresponding variable y_{ik}, or the slack of the corresponding dual constraint, is strictly negative. For each cell with existing negative reduced costs, we solve a column generation subproblem to search for a column with the minimum negative reduced cost. The column is used to enlarge M1-EFDA-MAS, which is then reoptimized. If, on the other hand, all reduced costs are nonnegative, then the optimal solution to M1-EFDA-MAS is also optimal to M1-EFDA-LP.

Because of the small size of the sets $\mathcal{K}'_i$, M1-EFDA-MAS is usually easy to solve. To find the reduced costs, we need the optimal dual solution to M1-EFDA-MAS. Given a dual optimal solution (π, μ) to M1-EFDA-MAS, where $\mu = \{\mu_i : i \in \mathcal{I}\}$ are the dual variables associated with (3.46c) and $\pi = \{\pi_j : j \in \mathcal{J}\}$ are the dual variables associated with (3.46b), then, using LP-duality, the reduced cost of y_{ik} is

$$\bar{c}_{ik} = L_{ik} - \sum_{j\in\mathcal{J}_i} \pi_j a_{ijk} - \mu_i \ . \tag{3.47}$$

Clearly, for cell i, there exists a y_{ik}-variable with a negative reduced cost if and only if the minimum of $\bar{c}_{ik}, k \in \mathcal{K}_i$, is less than zero. Let $\{q_k, k \in \mathcal{K}_i\}$ be the set of decision variables that indicate whether column k is to be added to the restricted master problems or not. The column generation subproblem for cell i can be formulated as follows.

$$\text{[M1-EFDA-SUB}_i\text{]} \quad \sum_{k\in\mathcal{K}_i\setminus\mathcal{K}'_i} (L_{ik} - \sum_{j\in\mathcal{J}_i} \pi_j a_{ijk}) q_k \quad \rightarrow \min \tag{3.48a}$$

$$\text{s. t.} \sum_{k\in\mathcal{K}_i\setminus\mathcal{K}'_i} q_k \leq 1 \tag{3.48b}$$

$$q_k \in \{0,1\} \qquad k \in \mathcal{K}_i \setminus \mathcal{K}'_i \tag{3.48c}$$

If $k^* = \arg \min_{k \in \mathcal{K} \setminus \mathcal{K}'_i} \{L_{ik} - \sum_{j \in \mathcal{J}_i} \pi_j a_{ijk}\}$, then the optimal solution to M1-EFDA-SUB$_i$ is

$$q_k = \begin{cases} 1 & \text{if } k = k^*, \\ 0 & \text{otherwise.} \end{cases}$$

For cell i, if the optimal objective function value of M1-EFDA-SUB$_i$ is less than μ_i, then variable q_{k^*} is used to enlarge M1-EFDA-MAS, i.e., set $\mathcal{K}'_i$ gets element k^*. After solving $|\mathcal{I}|$ subproblems, at most $|\mathcal{I}|$ variables will be added to M1-EFDA-MAS in one iteration. The restricted master problem is then re-optimized, and the algorithm proceeds to the next iteration. If, however, $\bar{c}_{ik} \geq 0, \forall i \in \mathcal{I}, \forall k \in \mathcal{K}_i$, then the optimal solution to M1-EFDA-MAS is also optimal to M1-EFDA-LP. This means that the optimum of the LP relaxation equals the Lagragian dual optimum.

In the worst case, after a finite number of iterations, all the y-variables are added to M1-EFDA-MAS, which then becomes identical to M1-EFDA-LP. However, usually only a few of all columns are necessary before M1-EFDA-LP is solved to optimality. For large-scale instances, this greatly reduces the computational effort for solving M1-EFDA-LP. Note also the similarities between M1-EFDA-SUB$_i$ and M1-EFIP-R1$_i$: $L_{i\bar{k}} = \sum_{k \leq \bar{k}} L^I_{ik}$ and $\sum_{j \in \mathcal{J}_i} \pi_j a_{ij\bar{k}} = \sum_{k \leq \bar{k}} \sum_{j \in \mathcal{J}_{ik}} \lambda_j, \forall \bar{k} \in \mathcal{K}_i$.

In the first iteration of our algorithm, we need to initialize the sets $\mathcal{K}'_i, \forall i \in \mathcal{I}$. These initial sets should ensure that M1-EFDA-MAS is feasible. This can be easily done using the solution of gain-based pilot power (see Section 3.4.2). For cell i, we set $y_{ik} = 1$ if and only if

$$k = \arg \min_{\substack{k \in \mathcal{K}_i: \\ L_{ik} \geq P_i^G}} \{L_{ik}\} \ , \tag{3.49}$$

where P_i^G is defined by (3.20). Let k_i denote the index of the CPICH power level, for which $y_{ik_i} = 1$. Then, in the first iteration of our algorithm, k_i is the only element in the set $\mathcal{K}'_i$. Doing so for all cells yields the initial master problem, which is clearly feasible (i.e., every bin is covered by at least one cell).

3.7.2 An Iterative Rounding Procedure

In Section 3.7.1, a column generation method has been presented for solving the LP-relaxation of M1-EFDA. To ensure integer optimality, a branch-and-bound scheme, which embeds the column generation algorithm into the enumeration tree, is necessary. This would require, however, very long computing time for large networks. We therefore use an iterative rounding procedure which allows us to generate a near-optimal solution.

In one iteration, the procedure rounds one fractional-valued variable in the optimal solution of the LP-relaxation to one. The rounding is based on the optimal solution of the LP-relaxation. Let $y^*_{ik}, \forall i \in \mathcal{I}, k \in \mathcal{K}'_i$, be the optimal solution to M1-EFDA-LP. The rounding procedure chooses the variable with the largest value among all fractional-valued variables. Let $y^*_{i^*k^*}$ denote this variable, that is,

$$y^*_{i^*k^*} = \max_{\substack{i \in \mathcal{I}, \\ k \in \mathcal{K}_i^F}} y^*_{ik} \ , \tag{3.50}$$

where $\mathcal{K}_i^F = \{k \in \mathcal{K}'_i : 0 < y^*_{ik} < 1\}$. If the optimal solution to M1-EFDA-LP does not contain any fractional values, i.e., $\mathcal{K}_i^F$ is empty, this solution is also optimal to M1-EFDA. Otherwise, variable $y^*_{i^*k^*}$ is rounded to one, by adding a new constraint to M1-EFDA-LP, e.g.,

$$y_{i^*k^*} = 1 \ . \tag{3.51}$$

Algorithm I.4 Column generation embedded into an iterative rounding procedure

Input: $\mathcal{I}, \mathcal{J}, \{\mathcal{K}_i, i \in \mathcal{I}\}, \{L_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}, \{P_i^G, i \in \mathcal{I}\}$
Output: LB, P^*, y^*

```
1: {K'_i, i ∈ I} ⇐ initRestrMAS({P^G_i, i ∈ I}) // Initialize sets {K'_i, i ∈ I}
2: for iter = 1, ..., |I| do
3:     repeat
4:         found ⇐ false
5:         (y, μ, π) ⇐ solveRestrMAS({K'_i, i ∈ I})
6:         for ∀i ∈ I do
7:             for ∀k ∈ K_i do
8:                 c̄_ik ⇐ computeRedCost(μ, π) // Compute reduced cost
9:             end for
10:            k' ⇐ arg min_{k∈K_i} c̄_ik
11:            if c̄_ik' < 0 then
12:                K'_i ⇐ K'_i ∪ {k'}
13:                found ⇐ true
14:            end if
15:        end for
16:    until ¬found
17:    if iter == 1 then
18:        LB ⇐ objEFDALP(y)
19:    end if
20:    for ∀i ∈ I do
21:        K^F_i ⇐ {k ∈ K'_i : 0 < y_ik < 1}
22:    end for
23:    numFracVar ⇐ Σ_{i∈I} |K^F_i|
24:    if numFracVar > 0 then
25:        (i*, k*) ⇐ arg max_{i∈I, k∈K^F_i} y_ik
26:        addConstr("y_{i*k*} = 1") // Add a new constraint to M1-EFDA-LP
27:    else
28:        P* ⇐ objEFDALP(y)
29:        y* ⇐ y
30:        break
31:    end if
32: end for
```

Adding constraint (3.51) makes the current solution $y^*_{ik}, i \in \mathcal{I}, k \in \mathcal{K}'_i$, infeasible, because $y^*_{i^*k^*} < 1$. We need therefore to re-optimize M1-EFDA-MAS. In addition, re-optimization of M1-EFDA-MAS leads to new values of the dual variables, which, in turn, may result in negative reduced costs for some y-variables that are currently not present in the sets $\mathcal{K}'_i, i \in \mathcal{I}$. In other words, we may need to add new elements to the sets $\mathcal{K}'_i$, in order to solve M1-EFDA-LP with the new constraint (3.51) to optimality.

The iterative rounding procedure, which repeatedly applies column generation to a sequence of LPs, is summarized in Algorithm I.4. In the presented algorithm, the master problem (M1-EFDA-LP) is solved in lines 3–16, and the restricted master problem (M1-EFDA-MAS) is solved in line 5. Rounding is performed in lines 25-26. Function **objEFDALP**(y) stands for computing the objective function values for a given solution in M1-EFDA-LP and M1-EFDA, respectively. It can be easily realized that the procedure generates an integer solution within a finite number of iterations (at most $|\mathcal{I}|$ iterations). The algorithm input is

→ a set of cells $\mathcal{I}$,

→ a set of bins $\mathcal{J}$,

→ a set of power level indices $\{\mathcal{K}_i, i \in \mathcal{I}\}$,

→ power parameters $\{L_{ik}, i \in \mathcal{I}, k \in \mathcal{K}_i\}$,

and the output consists of the lower bound LB, an integer solution y^* and the corresponding objective function value P^*.

3.8 Numerical Studies

In this section, computational results obtained for seven test networks of various sizes are presented.The information about the networks as well as some network statistics are given in Appendix A. Test networks Net1-Net7 are used for numerical studies in this section. The parameter setting used in the numerical experiments is shown in Table A.3 in Appendix A. Since we assume that the networks are highly loaded, the DL load factor $\eta_i^{DL}, i \in \mathcal{I}$, in all our experiments is set to one.

All numerical experiments have been conducted on an HP ProLiant DL385 with two AMD Opteron Processor 285 (dual-core, 2.6 GHz, 1MB cache, 4GB RAM). For finding optimal solutions to MIP and LP problems, we have used commercial solver ILOG CPLEX [29]. The Lagrangian heuristic was implemented in C++, and column generation with the rounding procedure was implemented as a C-program that uses CPLEX Callable Library for interaction with the CPLEX solver. The choice of programming languages for the two approaches is mainly due to computational efficiency and memory utilization considerations. Ad hoc solutions (see Section 3.4) are not computationally costly and therefore the selected programming language and the implementation environment are not critical, e.g., they can be computed within a few seconds in Matlab even for our largest test network.

3.8.1 Numerical Experiments without Ensuring Smooth Handover and Discretizing the CPICH Power Range

In the first part of the study, we present numerical results without ensuring smooth handover in the preprocessing step and assuming that no set with discrete power levels is explicitly given. Table 3.1 shows the two ad hoc solutions for the uniform pilot power and the gain-based approaches discussed in Section 3.4. For each of the two approaches, the first column in the corresponding part of the table shows the total pilot power over the network, i.e., $|\mathcal{I}| \cdot P^U$ for the first approach and $\sum_{i \in \mathcal{I}} P_i^G$ for the second approach, and the average pilot power per cell is depicted in the next column. The overlap metric is computed as the percentage of the area with CPICH signals from two or more cells satisfying the RSCP and the E_c/I_0 requirements. These areas are important since they can be viewed as potential SHO zones. Having some overlap is desirable to ensure that users and the network can gain from SHO, although large amount of overlap may result in high pilot pollution and large amount of SHO overhead that may lead to negative SHO gain. Typically, approximately 30 % of users are assumed to be in SHO in static planning scenarios, meaning that the overlap should be slightly larger than this value. Note, however, that since the results were obtained for the worst-case interference scenario, the cell overlap is most likely larger when the solution is applied in practice. With this note we could assume a target coverage overlap of 30 % if the planning is done for a highly loaded network.

We observe that the overlap is very high when the uniform pilot power approach is applied. A significant improvement is achieved by the second approach, although the overlap is still a bit higher than desired. The only exception is Net1 where the overlap is 8.29 %, i.e., much below 30 %, which is due to the characteristics of the modeled terrain and weaker fading effects. Comparing the pilot power values obtained by the two approaches, we observe that the gain-based approach significantly outperforms the uniform pilot power. Among the test

Table 3.1: Ad hoc solutions for full coverage

Network	Uniform CPICH power approach			Gain-based CPICH power approach		
	Total power [W]	Average power [W]	Overlap [%]	Total power [W]	Average power [W]	Overlap [%]
Net1	71.23	1.19	27.64	37.27	0.62	8.29
Net2	182.73	2.54	69.70	104.45	1.45	36.13
Net3	224.84	2.50	62.92	135.49	1.50	37.26
Net4	271.56	2.51	66.62	171.34	1.59	40.99
Net5	307.88	2.39	62.15	217.12	1.68	43.79
Net6	365.98	2.47	72.51	161.82	1.16	41.01
Net7	285.91	2.04	65.11	143.32	1.02	37.38

Table 3.2: Optimal solutions obtained using CPLEX

Network	Total power	Average power	Overlap	CPU	Gap (in %) with solution of	
	[W]	[W]	[%]	time, [sec]	uniform power	gain-based power
Net1	33.53	0.56	3.64	0.05	112.4	11.2
Net2	91.38	1.27	23.77	3.08	100.0	14.3
Net3	118.59	1.32	28.38	95.57	89.6	14.3
Net4	149.00	1.38	30.00	1254.56	82.2	15.0
Net5	183.38	1.31	32.14	22322.52	67.9	18.4
Net6	-	-	-	-	-	-
Net7	-	-	-	-	-	-

Table 3.3: Solutions obtained by the Lagrangian heuristic

Network	Integer solution			Lower	Gap	CPU	Ref. CPU
	Total, [W]	Average, [W]	Overlap, [%]	bound, [W]	[%]	time, [sec]	time, [sec]
Net1	33.53	0.56	3.64	33.50	0.11	1.06	0.03
Net2	92.02	1.28	24.54	89.98	2.27	6.06	2.28
Net3	120.29	1.34	29.40	115.38	4.26	20.65	18.23
Net4	150.99	1.40	30.92	144.55	4.46	29.91	24.68
Net5	186.35	1.44	33.24	177.33	5.08	66.60	62.13
Net6	129.34	0.92	26.02	123.03	5.12	119.57	664.35
Net7	119.39	0.85	28.72	109.76	8.76	734.96	24630.86

networks, the minimum improvement (29.7 %) has been obtained for Net5, and the maximum improvement (50.0 %) has been found for Net7.

Table 3.2 presents integer optimal solutions to the pilot power optimization problem M1 for the first five test networks. For the last two networks, the solver was not able to find an optimal solution within a reasonable amount of time since the computing time increases exponentially with the problem size. The presented results were obtained for formulation M1-EFIP, although there is no substantial difference between the computing times for M1-EFIP and M1-EFDA. In general, for small networks M1-EFDA performed slightly better but its computing time increases more rapidly with the problem size as compared to M1-EFIP. For the cell-bin formulation (M1-CBF) the solver was not able to solve the problem within one hour even for the smallest network Net1, which demonstrates the efficiency of the enhanced formulations. The two last columns in Table 3.2 present the gaps between the optimal solution and the uniform and the gain-based pilot power solutions, respectively. We observe that although the gain-based pilot power solutions are significantly closer to the optimal solutions, there is still some space for improvement.

Tables 3.3 and 3.4 present solutions obtained by the Lagrangian heuristic and the column generation approach, respectively. For each of the solution in the tables, we show the

Table 3.4: Solutions obtained by the column generation approach

Network	Integer solution			Lower	Gap	CPU	Ref. CPU
	Total, [W]	Average , [W]	Overlap, [%]	bound, [W]	[%]	time, [sec]	time, [sec]
Net1	33.70	0.56	4.00	33.50	0.61	0.05	0.01
Net2	92.11	1.28	24.74	90.23	2.09	1.20	1.77
Net3	120.69	1.34	29.59	116.61	3.50	10.30	18.73
Net4	149.80	1.39	30.09	146.60	2.18	14.78	20.31
Net5	187.02	1.45	33.65	179.35	4.28	117.18	57.13
Net6	127.71	0.86	24.16	124.37	2.71	1043.27	1334.35
Net7	117.73	0.84	27.04	111.73	5.17	40981.34	39025.53

total and average pilot power, the coverage overlap percentage, the lower bound, and the gap between the found integer solution (upper bound) and the lower bound. The two last columns of the tables show the CPU time for obtaining the results and the CPU time ("Ref. CPU time") spent by CPLEX to obtain a solution of the same quality in optimality gap. Comparing the CPU times, we observe that the Lagrangian heuristic significantly outperforms the CPLEX engine (under the default configuration and parameter setting) when it comes to large networks, e.g., Net6 and Net7 in our experiments. For smaller networks, the computational times for getting near-optimal solutions are comparable.

Tables 3.3 and 3.4 demonstrate that the integer solutions obtained by the two approaches are comparable. For networks Net1-Net3 and Net5, better upper bounds were found by the Lagrangian heuristic, whilst for networks Net4, Net6, and Net7 the approach based on column generation found better upper bounds. The second approach generated also the best lower bounds (since the LP bound from column generation is actually the optimum to the Lagrangian dual) and achieved the best gaps for all the test networks. For the large networks, this is, however, at a cost of significantly longer computation times. Figure 3.4 visualizes the total pilot power of the solutions for all seven test networks.

Next we provide a more detailed analysis of the obtained results for Net1, Net6, and Net7. These networks use realistic data. In Figures 3.5(a) and 3.5(b) we show the cumulative distribution of the best-server RSCP and CIR levels, respectively, in the solutions of uniform, gain-based, and optimized pilot power for the three test networks. In every bin, the best server is defined as the cell from which the strongest signal is received, i.e., with the highest RSCP or E_c/I_0 ratio. The two ratios are equal when the denominator in (3.1) is the same for all cells. For optimized pilot power we use the solutions presented in Table 3.3. In Figure 3.5(a) we observe that the distribution curves for the optimized power are shifted by 1–2 dB to the left compared to the uniform power solutions. The curves' tails, however, are almost not affected because of the full coverage requirement, i.e., because there exist some bins that even in the optimized solution require a pilot power level close to that of the uniform solution. Analyzing the distribution curves in Figure 3.5(b), we can conclude that for the optimized pilot power the CIR levels in a majority of bins (70–80 %) are within the interval $[-20, -14]$ dB, while in the solutions of uniform power about the same number of bins have a CIR of -13 dB or higher. The distribution curves of the gain-based solutions are close to the optimized power solutions.

Figure 3.6 addresses the issue of pilot pollution. It shows the cumulative distributions of the differences between the first and the second, the third, and the fourth strongest signal in each of the three solutions for Net6. Usually, it is assumed that the fourth CPICH signal is the first polluting CPICH (for example, in [46]). This is because it has been shown that the maximum practical active set size in WCDMA networks is three (see, for example, [19]). In rich multi-path environments, even three CPICH signals can cause pollution due to insufficient number of RAKE fingers [46]. Observe that area having four CPICH signals has been significantly reduced in the gain-based power solution and is almost completely avoided

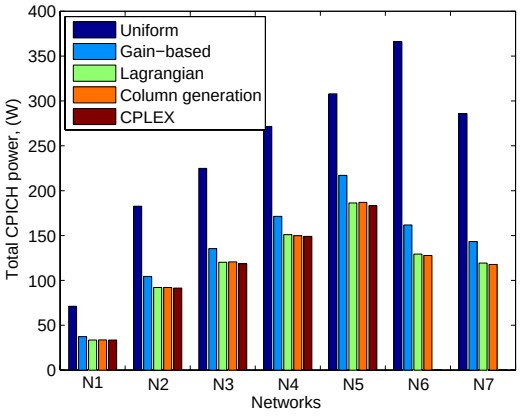
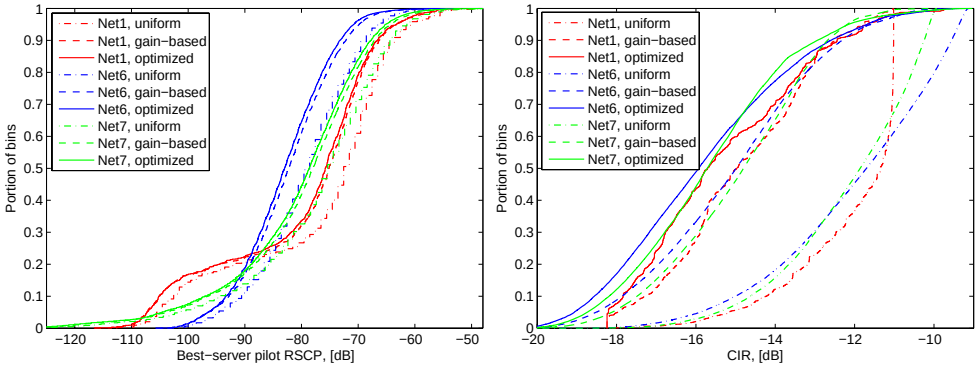


Figure 3.4: Total CPICH power in solutions obtained by different approaches.



(a) Best-server RSCP distribution

(b) Best-server CIR distribution

Figure 3.5: CPICH signal strength characteristics in the uniform, gain-based, and optimized pilot power solutions for Net1, Net6, and Net7.

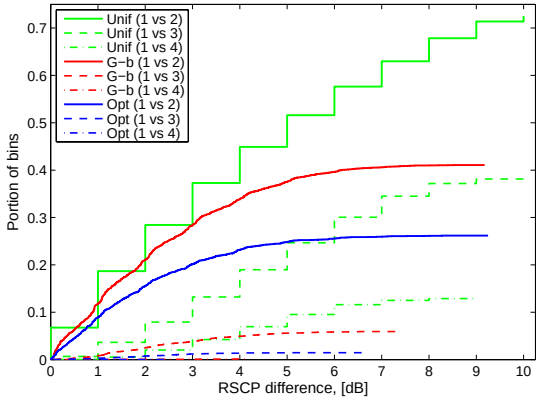
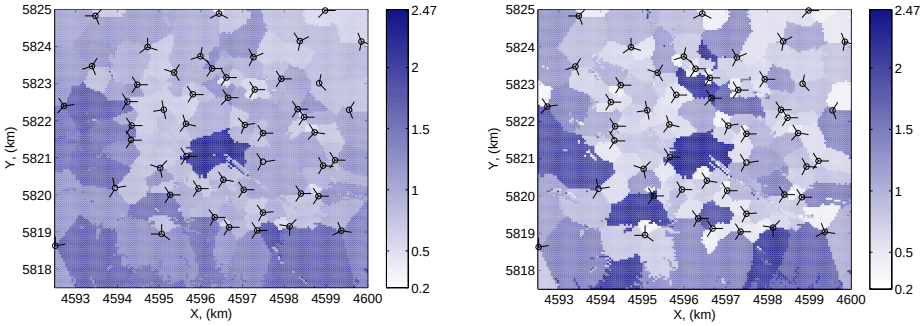
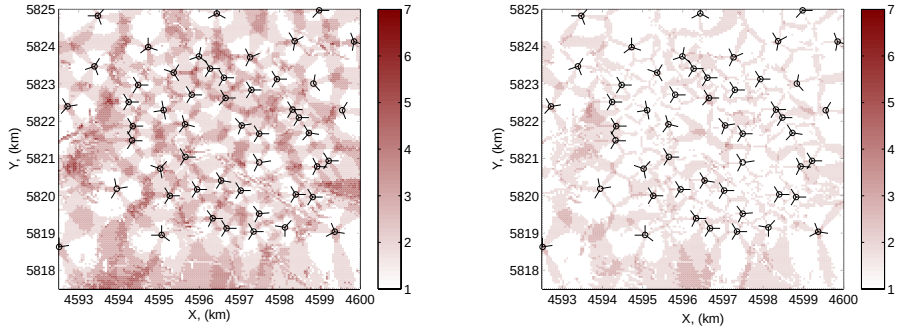


Figure 3.6: CPICH RSCP difference between the serving cell and the second, third, and fourth covering signals.

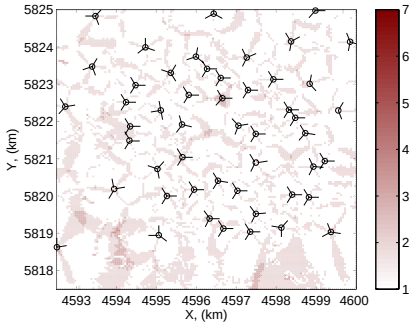


(a) Solution of gain-based CPICH power (b) Solution of optimized CPICH power

Figure 3.7: Best-server CPICH power in two solutions for Net6, [W].

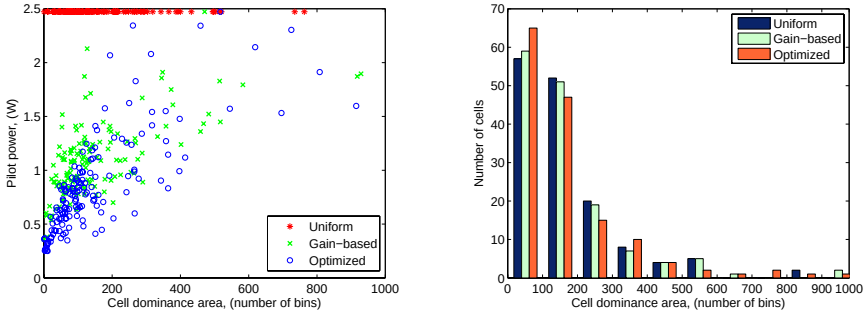


(a) Solution of uniform CPICH power (b) Solution of gain-based CPICH power



(c) Solution of optimized CPICH power

Figure 3.8: Coverage statistics in the ad hoc and optimized CPICH power solutions for Net6.



(a) CPICH power vs cell dominance area (b) Cell dominance area histogram

Figure 3.9: Cell size statistics in the ad hoc and optimized CPICH power solutions for Net6.

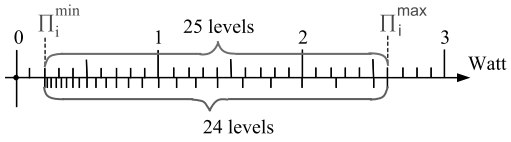


Figure 3.10: CPICH power range discretization on linear and logarithmic scales (Net2-Net7).

in the optimized pilot power solution. In the uniform power solution, in about 10 % of the area the fourth CPICH signal can be heard and is quite strong compared with the best signal (the RSCP difference is 5 dB). The area where the third CPICH signal can be heard has been reduced by more than a factor of two in the gain-based power solution compared to the uniform power solution, and is about 1 % in the optimized power solution.

Figures 3.7(a) and 3.7(b) demonstrate the gain-based power solution and the optimized power solution, respectively, for the city of Berlin (Net6). For the same network, the coverage overlap statistics for the uniform, gain-based, and optimized power solutions are shown in Figures 3.8(a), 3.8(b), and 3.8(c), respectively, where we for each bin depict the number of cells satisfying the RSCP and the E_c/I_0 thresholds. The maximum number of cells covering a bin in the case of uniform pilot power is seven, and it is also very common with three or four covering cells. The coverage overlap significantly reduces in the two other solutions where the number of covering cells of any bin does not exceed four.

Figure 3.9(a) shows the CPICH power level and the size of the dominance area of each cell. The cell dominance area is defined as the area where the cell has the strongest CPICH RSCP among all covering cells, i.e., is the best server. We observe that in the optimized power solution most of the cells have CPICH power below 1 W and are characterized by the cell dominance area spanning less than 200 bins (i.e., approximately 700 m × 700 m). Interestingly, the correlation between the size of cell dominance area and the CPICH power beyond this region is much weaker, i.e., cells with large dominance areas are not necessarily those with high CPICH power. Exploring Figure 3.7, we note that large cells mostly occur along the borders of the area of interest, e.g., at the bottom of the figure, where the sites are sparsely located due to a low amount of predicted traffic in those areas. More attention require the cells that have the highest CPICH power levels and at the same time have average cell sizes, because they reveal the necessity of reconfiguring their own sites and/or their neighboring sites (e.g., changing antenna azimuth and/or tilt to reduce interference in the critical areas). The histograms of the cell dominance area sizes in the three solutions for Net6 are shown in Figure 3.9(b) where the size is measured in the number of bins.

3.8.2 Numerical Solutions Obtained by Discretizing the CPICH Power Range

The next study focuses on approximations obtained through discretizing the continuous range of the CPICH power on linear and logarithmic scales. When the set of possible CPICH power levels is not explicitly given to the network planner, but is constructed from computed P_{ij}-values, the problem size tends to be much larger than as if we were given discrete power levels taken from a range with a certain (not too small) step. The latter would allow us to solve the problem faster, but, on the other hand, the solution quality can be affected. Here, we compute solutions for ranges discretized on linear and logarithmic scales, and study how the results and the computational time are affected by the scale choice. The study was motivated by the fact that the distribution of P_{ij}-values is not uniform, and therefore choosing the logarithmic scale seems to be more natural. On the other hand, having larger discretization step for high P_{ij}-values may result in higher cell CPICH power levels, if those bins define the CPICH power of some cells, and thus may lead to a larger gap between the

Table 3.5: Optimal and near-optimal solutions for a set of power levels obtained by discretization on linear scale

Network	CPLEX			Lagrangian heuristic			
	Total power, [W]	Average power, [W]	CPU time, [sec]	Total power, [W]	Average power, [W]	CPU time, [sec]	Ref. CPU time, [sec]
Net1	36.3	0.61	0.01	36.3	0.61	0.07	0.01
Net2	94.5	1.31	1.64	95.0	1.32	1.05	0.86
Net3	122.6	1.36	14.93	123.5	1.37	3.38	7.32
Net4	154.1	1.43	24.02	156.4	1.45	5.47	10.04
Net5	189.5	1.47	253.58	191.8	1.49	8.99	16.64
Net6	134.4*	0.91	14744.30	136.2	0.92	17.31	51.81
Net7	122.8**	0.88	41592.76	124.5	0.89	56.83	132.80

* Near-optimal solution with optimality gap of 1 %
** Near-optimal solution with optimality gap of 2 %

Table 3.6: Optimal and near-optimal solutions for a set of power levels obtained from discretization on logarithmic scale

Network	CPLEX			Lagrangian heuristic			
	Total power, [W]	Average power, [W]	CPU time, [sec]	Total power, [W]	Average power, [W]	CPU time, [sec]	Ref. CPU time, [sec]
Net1	35.60	0.59	0.02	35.60	0.59	0.09	0.02
Net2	96.20	1.34	0.74	96.61	1.34	1.07	0.64
Net3	125.61	1.39	8.21	126.24	1.40	3.21	4.18
Net4	157.00	1.45	15.42	158.62	1.48	4.90	5.17
Net5	193.29	1.50	34.03	194.84	1.51	7.68	10.19
Net6	134.28*	0.91	2915.09	135.79	0.92	17.39	91.97
Net7	122.77**	0.88	11633.06	125.49	0.90	59.22	120.11

* Near-optimal solution with optimality gap of 1 %
** Near-optimal solution with optimality gap of 2 %

approximation and the optimal solution.

Tables 3.5 and 3.6 present results for linear and logarithmic scales, respectively. Consider a continuous pilot power range of $[\Pi_i^{min}, \Pi_i^{max}]$. Assume steps of 0.1 W and 0.5 dB for the corresponding scales. Then, with the parameter setting given in Table A.3 in Appendix A, the discrete sets of power levels covering the given range, for example, for Net2-Net7 are $\{0.2, 0.3, 0.4, \ldots, 2.6\}$ and $\{-7, -6.5, \ldots, 3.5, 4, 4.15\}$ in linear and logarithmic scale, respectively. This is illustrated in Figure 3.10. The first set contains 25 levels, and the second set has 24 levels. We applied the Lagrangian heuristic to find solutions to the problem with pre-discretized power levels. The optimal solutions were found by Cplex.

We observe that the optimal solutions for the discrete power levels sampled on the linear scale outperform those for the logarithmic scale in all solutions, except Net1. This is in line with our observations that large pilot power levels are few but at the same time critical for the solution quality. This does not hold for Net1, partly because of the smaller range ($[0.2, 2.0]$ Watt) and partly because of less variation in the P_{ij}-values due to a smoother terrain. Although the difference in total power is at most 4.5 %, the difference in computational times is tremendous for most of the test networks, which is explained by that bounding and fathoming is easier with a smaller power discretization step which is the case for pilot power levels below 0.9 Watt sampled on logarithmic scale. (Note that in the optimal solutions, the pilot power in most cells is in fact below 0.9 W, as can be seen from Figure 3.9(a).) Comparing the optimal solutions in Table 3.5 to those in Table 3.2, we find that the gap between the linear-scale approximation and optimum is about 3–8 %.

Table 3.7: Ad hoc solutions with smooth handover

Network	Uniform CPICH power solution				Gain-based CPICH power solution			
	Total power, [W]	Average power, [W]	Overlap [%]	Incr. [%]	Total power, [W]	Average power, [W]	Overlap [%]	Incr. [%]
Net1	104.98	1.75	37.09	47.4	56.12	0.94	20.95	50.6
Net2	219.03	3.04	77.66	19.9	142.70	1.98	58.91	36.6
Net3	347.68	3.86	80.58	54.6	186.91	2.08	57.55	37.9
Net4	398.47	3.69	80.56	46.7	226.81	2.10	58.65	32.4
Net5	441.97	3.43	75.95	43.5	280.71	2.18	59.06	29.3
Net6	486.52	3.29	79.36	32.9	216.39	1.46	56.17	33.7
Net7	432.97	3.09	76.12	51.4	189.97	1.36	50.08	32.5

Table 3.8: Lagrangian heuristic solutions with smooth handover

Network	Integer solution				Lower bound [%]	Gap [%]
	Total power, [W]	Average power, [W]	Overlap [%]	Incr. [W]		
Net1	52.43	0.87	19.49	56.4	52.43	0.10
Net2	131.32	1.82	54.04	42.7	128.76	1.98
Net3	162.34	1.80	49.75	34.9	156.59	3.67
Net4	205.02	1.90	52.05	35.8	194.54	5.39
Net5	245.78	1.91	52.11	31.9	233.64	5.20
Net6	184.94	1.25	48.44	42.9	175.93	5.12
Net7	164.27	1.17	45.92	37.6	150.61	9.07

3.8.3 Numerical Results for Smooth Handover

Extending cell coverage may be needed in such scenarios like Net1 where the coverage overlap in the optimized power solution is too small. We therefore study how coverage extension in all cells ensured in the preprocessing step by adjusting the P_{ij}-values can affect the total pilot power in the network. We increase the cell overlap to ensure smooth handover as described in Section 3.2 and find the uniform, gain-based, and optimized power solutions as we did in Section 3.8.1. The first two solutions are presented in Table 3.7 and the optimized power solutions obtained by the Lagrangian heuristic are presented in Table 3.8. Computational times for the Lagrangian heuristic solutions are very similar to those shown in Table 3.3 and therefore are not presented in this section. In both tables we compute the relative increase (in %) in the total power due to ensuring smooth handover. This increase can be viewed as the price of extending cell coverage. The total CPICH power has increased by 19.9–56.4 %, which shows that extending the coverage is quite expensive from the resource point of view. However, the optimized solutions for smooth handover still outperform the uniform power solutions without the smooth handover constraints (presented in Table 3.3).

3.9 Discussion and Conclusions

In this chapter, we have discussed pilot power optimization for full coverage of the service area. We have presented a mixed-integer programming formulation and derived two enhanced integer programming formulations. Furthermore, we have developed two heuristic algorithms. One utilizes Lagrangian relaxation, and the second is based on column generation and rounding. The solutions for different test networks obtained by these algorithms have been analyzed and compared to those obtained by ad hoc approaches and the optimal solutions obtained by Cplex. The computational results show that the proposed algorithms yield significant improvements as compared to ad hoc solutions to this problem. The pilot power savings are up to 25 % compared to the gain-based solutions and up to 67 % compared

to the uniform pilot power solutions. Another advantage of the presented algorithms is that they can give a feasible solution of high quality even for large networks, on which a standard solver fails, with a reasonable amount of computational effort. From a practical point of view, the quality of the solutions found by the algorithms is sufficiently high for the purpose of network planning, because of the uncertainty in the network data (in particular, the power gain values).

When analyzing coverage results and the CPICH power distribution among the cells, it is important to note that cells with a pilot power level equal or close to Π_i^{min} (for corresponding i) maybe considered for removal ("switching off"), may need reconfiguration, or may require reconfiguration of strongly interfering neighboring cells. To make a decision in such a situation, a more detailed analysis taking also into account traffic demand is needed.

We have numerically studied the problem with a pre-discretized set of pilot power levels and observed that discretization on linear scale results in solutions of higher quality, whereas discretization on logarithmic scale can be helpful in obtaining fast solutions. Our numerical experiments for smooth handover, which led to extended cell coverage due to adjusting power gains in the preprocessing step, demonstrated that even a slight cell coverage extension may be very costly from the resource point of view and therefore needs careful planning.

Chapter 4

Pilot Power Optimization for Partial Coverage

Pilot power optimization always involves a trade-off between pilot power consumption and service area coverage. This trade-off has been already studied in the previous chapter, but the focus was minimization of pilot power consumption under the requirement of full coverage of the service area. Here, the basic pilot power optimization model is extended to allow solutions satisfying a given minimum coverage degree, where the coverage degree can be either a relative size of the service area or relative traffic demand. We present an integer programming formulation for the extended problem and propose modifications to the algorithms presented in the previous chapter to enable near-optimal solutions within a reasonable amount of time. Numerical experiments for realistic planning scenarios are designed for studying the effect of coverage degree on the total amount of CPICH power in the network.

4.1 Motivation

Full coverage of the service area is a desired network property, but in real-life networks providing full CPICH coverage is usually very expensive from both economic and resource consumption point of view. Therefore, in practice, a guaranteed coverage level of 95-98 % would be sufficient for any UMTS network. In [38, 44], for example, coverage probability of 90–95 % is mentioned as a typical assumption when planning coverage and optimizing radio resources for UMTS.

In system management, there has been practically proved a phenomenon known as *Pareto's Principle*, or the 80/20 rule, which says that 20 % of the efforts always are responsible for 80 % of the results (see Figure 4.1). This phenomenon suggests that a slight decrease in the coverage degree (*goal*) may enable considerable reductions in the pilot power and network resource utilization in general (*efforts*). In this chapter, we intend to investigate if this rule also works for our model.

4.2 System Model and Optimization Problem

The system model for the CPICH optimization problem presented in this section extends the system model in Section 3.1. We use $\beta \in (0, 1]$ to denote the required degree of traffic coverage. Note that $\beta = 1.0$ corresponds to full coverage, i.e., the CPICH optimization problem under the full coverage requirement is a special case of the problem discussed here. Thus, the problem considered in this chapter is also $\mathcal{NP}$-hard in general.

To estimate coverage of traffic in the network, we introduce a new parameter, *traffic demand*. One possible interpretation of this parameter is the average number of active users requesting a specific service in a bin. Another possibility is to consider the number of active

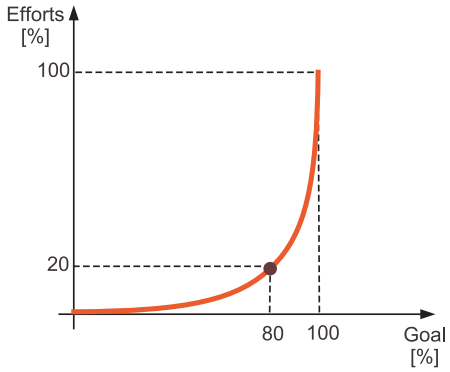


Figure 4.1: Pareto's principle (80/20 rule).

users of a specific service at a specific time, e.g., busy hour call attempts. Instead of the traffic demand for a specific service, it is also possible to consider traffic demand for a number of services, taken as a weighted average with different weighting coefficients for different services.

We use d_j to denote the traffic demand in bin j, and D to denote the total traffic demand over the network ($D = \sum_{j \in \mathcal{J}} d_j$). In a real network, user traffic demand is dynamic and non-uniformly distributed over the network. In static and short-term dynamic simulations, traffic demand can be modeled by a snapshot. The required coverage degree models the relative amount of traffic demand located in the area covered by CPICH signals and is denoted by β ($0 < \beta \leq 1$).

The problem of CPICH optimization for partial coverage (denoted by M2) is stated below.

Objective

- Find a vector $P^{CPICH} = \{P_i^{CPICH}, i \in \mathcal{I}\}$ that minimizes the total amount of pilot power in the network, i.e., $\sum_{i \in \mathcal{I}} P_i^{CPICH} \rightarrow \min$.

Constraints

- The CPICH coverage degree of at least β is to be achieved.
- The pilot power in any cell i is within a given interval, i.e., $\Pi_i^{min} \leq P_i^{CPICH} \leq \Pi_i^{max}$.

For a uniform traffic distribution, the definition of coverage degree is rather clear straightforward. We define the coverage degree as the relative size of the entire service area which is actually covered. In this case, to measure the coverage degree, we find the portion of bins for which (3.1) holds. Let us refer to this type of coverage as *service area coverage*.

In a real network the traffic distribution is never uniform, which means that if we measure the coverage degree without considering traffic distribution, there is a risk that the solution will suffer from coverage loss in some area with high traffic intensity. To avoid this possible shortcoming of the model, we use a new measure of coverage, called *traffic coverage degree*, which is defined as the ratio between the sum of the amount of traffic demand in bins where (3.1) holds and the total traffic demand in the area.

4.3 Ad Hoc Solutions

Ad hoc solutions for the CPICH optimization problem with partial coverage can be obtained by adopting the two ad hoc strategies presented in Sections 3.4.1 and 3.4.2.

4.3.1 Uniform Pilot Power

Let D^j denote the total amount of traffic demand in the covered bins if the uniform pilot power level equals P^j, i.e.,

$$D^j = \sum_{l \in \mathcal{J}: P^l \leq P^j} d_l \ , \tag{4.1}$$

where P^j is the minimum CPICH power in a cell defined by (3.15), and d_l is the traffic demand in bin l.

To guarantee a given degree of traffic coverage, the total traffic demand from all the covered bins has to be at least βD, i.e., if $P^U = P^{j^*}$ for some bin j^*, then $D^{j^*} \geq \beta D$ must hold.

A straightforward approach is to sort the bins in ascending order with respect to P^j. The first element in the sorted sequence, for which the corresponding value of D^j is no less than βD, yields the minimum uniform pilot power that satisfies the coverage requirement. Mathematically, the solution can be formalized as follows,

$$P^U = \min_{j \in \mathcal{J}: D^j \geq \beta D} P^j \ . \tag{4.2}$$

4.3.2 Gain-based Pilot Power

There are several ways to obtain a power-minimization heuristic for the case of partial coverage by adapting the gain-based approach presented in Section 3.4.2. One such heuristic is as follows. For every bin, we find the maximum power gain among all cells, i.e., $g_{c(j)j}$, which is used to sort the bins in a descending order. The pilot power solution is then determined by including sufficiently many bins in the sorted sequence such that the total demand in these bins is at least βD.

4.4 Integer Programming Formulations

The mathematical programming models presented in this section are the extensions to the two formulations described in Section 3.5.2. In both formulations we use a set of binary variables $z_j, j \in \mathcal{J}$, defined as follows,

$$z_j = \begin{cases} 1 & \text{if bin } j \text{ is covered by at least one cell,} \\ 0 & \text{otherwise.} \end{cases}$$

The new set of variables is used to define a subset of bins that fulfils the requirement of partial coverage. Note that $z_j, j \in \mathcal{J}$, are all ones when full coverage is required.

To model the traffic coverage degree requirement, the following constraint needs to be added,

$$\sum_{j \in \mathcal{J}} d_j z_j \geq \beta D \ . \tag{4.3}$$

The next constraint models the service area coverage requirement,

$$\sum_{j \in \mathcal{J}} z_j \geq \beta |\mathcal{J}| \ . \tag{4.4}$$

Observe that (4.4) is a special case of (4.3). In particular, the two are equivalent in the case of uniform traffic distribution. Therefore, constraint (4.3) can be viewed as a generalized partial coverage requirement constraint.

4.4.1 A Formulation Based on Direct Assignment

Formulation M1-EFDA is considered as a base for the formulation derived in this section. To allow partial-coverage solutions, the right-hand side of constraints (3.22b) needs to be changed in the following way,

$$\sum_{i\in\mathcal{I}_j}\sum_{k\in\mathcal{K}_i} a_{ijk}y_{ik} \geq z_j \qquad j \in \mathcal{J}\ . \tag{4.5}$$

Adding also constraint (4.3) gives us the following formulation for the pilot power optimization problem with partial coverage.

$$\text{[M2-EFDA]}\quad P^* = \min \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L_{ik}y_{ik} \tag{4.6a}$$

$$\text{s.t.}\quad \sum_{j\in\mathcal{J}} d_j z_j \geq \beta D \tag{4.6b}$$

$$\sum_{i\in\mathcal{I}_j}\sum_{k\in\mathcal{K}_i} a_{ijk}y_{ik} \geq z_j \qquad j \in \mathcal{J} \tag{4.6c}$$

$$\sum_{k\in\mathcal{K}_i} y_{ik} \leq 1 \qquad i \in \mathcal{I} \tag{4.6d}$$

$$y_{ik} \in \{0,1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{4.6e}$$

$$z_j \in \{0,1\} \qquad j \in \mathcal{J} \tag{4.6f}$$

4.4.2 A Formulation Based on Incremental Power

The second formulation of the pilot power optimization problem with partial coverage can be derived from M1-EFIP. The formulation is presented below. Note that sets $\mathcal{K}_i, i \in \mathcal{I}$, contain sorted indices (this is not required in M2-EFDA).

$$\text{[M2-EFIP]}\quad P^* = \min \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} L^I_{ik}x_{ik} \tag{4.7a}$$

$$\text{s.t.}\quad \sum_{j\in\mathcal{J}} d_j z_j \geq \beta D \tag{4.7b}$$

$$\sum_{i\in\mathcal{I}_j} x_{i\kappa(i,j)} \geq z_j \qquad j \in \mathcal{J} \tag{4.7c}$$

$$x_{i(k-1)} \geq x_{ik} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \setminus \{1\} \tag{4.7d}$$

$$x_{ik} \in \{0,1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{4.7e}$$

$$z_j \in \{0,1\} \qquad j \in \mathcal{J} \tag{4.7f}$$

4.5 A Solution Approach Based on Lagrangian Relaxation

To solve the pilot power optimization problem with partial coverage, we can use the Lagrangian-based solution approach presented in Section 3.6, but with some modifications described below. The modified algorithm can be applied to both formulations presented in Section 4.4, i.e., M2-EFDA and M2-EFIP. We use M2-EFIP but to demonstrate the solution approach.

The z-variables in M2-EFIP in constraints (4.7c) imply a modification of the Lagrangian function. The new function is presented below,

$$LF(x,z,\lambda) = \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i}(L^I_{ik} - \sum_{j\in\mathcal{J}_{ik}}\lambda_j)x_{ik} + \sum_{j\in\mathcal{J}}\lambda_j z_j\ . \tag{4.8}$$

In M1-EFIP-R1 (see Section 3.6.2), the objective function (3.40) depends only on the x-variables that appear in the first part of Lagrangian function. For a given set of Lagrangian coefficients λ_j, the second part of the function $LF(x, \lambda)$ is constant. In the new Lagrangian function (4.8), the second part depends on the set of binary variables z_j, which means that the new relaxed problem decomposes into $|\mathcal{I}| + 1$ subproblems, i.e., we solve $|\mathcal{I}|$ Lagrangian subproblems for the x-variables (one for each cell), that is M1-EFIP-R1$_i, i \in \mathcal{I}$, and one subproblem for z-variables. Thus, lines (11) and (12) in Algorithm I.1 need to be replaced as shown below.

$$\begin{aligned} &11\text{: } (\mathrm{x}, \mathrm{z}) \Leftarrow \textbf{solveLagrRel}(\lambda) \\ &12\text{: } lb \Leftarrow LF(\mathrm{x}, \mathrm{z}, \lambda) \end{aligned}$$

The subproblem for z-variables is formulated as follows,

$$\begin{aligned} \text{[M2-EFIP-R2]} \quad & \sum_{j \in \mathcal{J}} \lambda_j z_j \quad \rightarrow \min \\ \text{s.t.} & \sum_{j \in \mathcal{J}} d_j z_j \geq \beta D \\ & z_j \in \{0, 1\} \qquad\qquad \forall j \in \mathcal{J} \end{aligned}$$

To find the optimal values of the z-variables that indicate the bins that must be covered to guarantee a given degree of traffic coverage, we solve subproblem M2-EFIP-R2 for a given set of Lagrange multipliers. Applying a simple substitution $z_j = 1 - \bar{z}_j$, problem M2-EFIP-R2 can be easily transformed to a binary knapsack problem. When the traffic demand values $d_j, j \in \mathcal{J}$, are integers or can be transformed to integers by using scaling, there is an efficient dynamic programming algorithm [56] that finds the optimal solution with $\mathcal{O}(|\mathcal{J}|\beta D)$ in time and $\mathcal{O}(\beta D)$ in space. When βD is a big number, the exact solution can be found by using integer programming. Also, scaling works well in practice, especially when the range of $d_j, j \in \mathcal{J}$ is not too large.

For large networks, finding an exact solution to M2-EFIP-R2 can be very costly in time. An approximation to the optimal solution can be found using a relaxation technique, e.g., LP-relaxation which tends to give good near-optimal solutions to M2-EFIP-R2 when variation in d_j values is small. This approach can be applied without ensuring the integrality of the traffic demand values d_j. The LP-relaxation is solved to optimality by sorting in descending order the ratios between the Lagrange multipliers and the traffic demand values, i.e., $\lambda_j/d_j, j \in \mathcal{J}$, and applying a greedy selection algorithm that picks the maximum number of sorted ratios from the head of the sequence such that the sum of the selected traffic demand values does not exceed $(1 - \beta)D$. The time complexity of solving the LP relaxation is determined by sorting. The LP-relaxation and heap sorting, which is $\mathcal{O}(|\mathcal{J}| \cdot \log |\mathcal{J}|)$, were used in our algorithm implementation for finding a solution to M2-EFIP-R2.

To adjust the algorithm in Section 3.6 to M2, a modification in the power adjustment heuristic is also needed. In both parts of the original power adjustment algorithm, we ensure the CPICH coverage in all bins. In its modified version, we ensure coverage of those bins only for which the corresponding z-variables are one. When reducing cell overlap (see Algorithm I.3, lines (8)-(13)), we skip those bins for which the z-variables are zero.

4.6 Numerical Results

In this section, solutions to the pilot power optimization problem with various traffic coverage degrees are examined. The computational results have been obtained for the following three test networks: Net1, Net6 (the city of Berlin), and Net7 (the city of Lisbon). All numerical experiments have been conducted on an HP ProLiant DL385 with two AMD Opteron Processor 285 (dual core, 2.6 GHz, 1MB cache, 4GB RAM).

The network statistics and the parameter setting used in our numerical experiments are presented in Appendix A. Figures A.2(a), A.2(b), and A.2(c) in Appendix A show the traffic demand distribution over the service area in test network Net1, Net6, and Net7, respectively. For all test networks, traffic demand is non-uniform and is given as a static load grid that contains the expected average number of users in each bin at one time instance. The traffic demand is represented by the speech-telephony service.

Tables 4.1, 4.2, and 4.3 present computational results for networks Net1, Net6, and Net7, respectively. Each of the tables shows three sets of solutions obtained for various coverage degrees: ad hoc solutions of uniform pilot power and gain-based pilot power discussed in Section 4.3 and solutions obtained by the Lagrangian heuristic discussed in Section 4.5. For each solution, we present the total amount of CPICH power and the overlap percentage. In addition, for the Lagrangian heuristic solutions we show the lower bound, duality gap, and computing times.

We observe that, in terms of power consumption, solutions obtained by the Lagrangian heuristic algorithm significantly outperform those of uniform pilot power, especially for higher levels of traffic coverage. In average, the optimized solutions give CPICH power savings of up to three times for full coverage and up to 25 % for 90 % traffic coverage ($\beta = 0.9$). Compared to the gain-based solutions, the improvement is less but still considerable as it varies from 11 to 25 %. Observe that the absolute improvement over the gain-based solutions in the amount of power does not vary a lot, whereas the variation is much larger with respect to the uniform solutions.

Comparing the duality gaps of solutions, we observe that the duality gap does not change in the same way for the three studied networks. Typically, it decreases when β goes down from 1.0, but then increases a bit when β approaches 0.9. This becomes especially obvious in Tables 4.1 and 4.3. One of the reasons is that for lower coverage degrees the CPICH power is more uniformly distributed among cells, which reduces the effectiveness of the coverage adjustment and the overlap-reduction algorithms, as they make use of sorted sequences of bins and cells, respectively.

As follows from the tables, computing times also depend on coverage level. In all cases, the times are reasonable considering the problem complexity and the problem size, especially for the two big networks Net6 or Net7. Obviously, the lower the required coverage degree, the less time is spent by the heuristic procedure because less coverage adjustment is needed to obtain feasible solutions. Therefore, the most time consuming part of the heuristic procedure becomes the one for reducing cell coverage overlap. This sometimes may cause longer computing times in comparison to solving the problem for full coverage, since, in addition to solving the knapsack problem, more CPICH power reduction steps may have to be performed. Examples of this behavior are the solutions for Net1 and Net6, where computing times for full coverage ($\beta = 1.0$) are shorter than those for most of the other coverage degree values.

Figure 4.2 shows the cumulative distributions of cell pilot power levels when traffic coverage degree β varies from 0.95 up to 1.0. Although for all values of β there are some cells that use relatively high CPICH power, the number of such cells decreases with decreasing β. For example, the best-server CPICH power is at most 1 W in 73 % of the area when $\beta = 1.0$, 80 % of the area when $\beta = 0.98$, and 88 % of the area when $\beta = 0.95$. Also, small variation in CPICH power is desirable for improving SHO gain [17, 27, 51].

Figures 4.3(a), 4.4, and 4.3(c) illustrate the total amount of CPICH power versus traffic coverage degree for uniform, gain-based, and optimized pilot power solutions. The lower bounds obtained by the Lagrangian heuristic are marked by dotted lines. We observe that for all test networks the curves follow well the Pareto's principle discussed in Section 4.1. That is, a significant reduction in CPICH power can be achieved by slightly decreasing the coverage degree.

We also observe that the difference between the uniform CPICH power solutions and the gain-based solutions decreases rapidly when β goes down. For Net6 and Net7, this difference

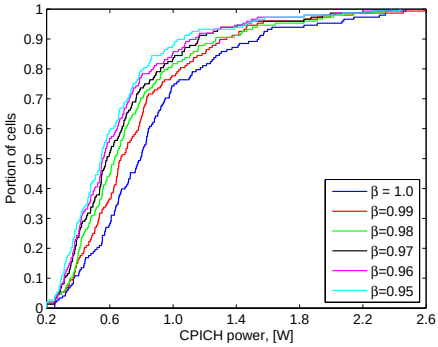
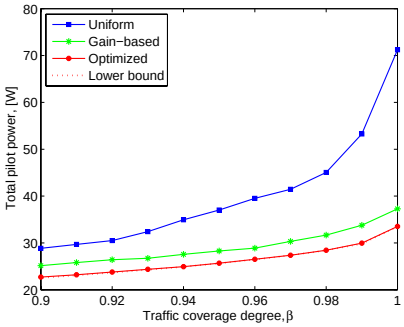
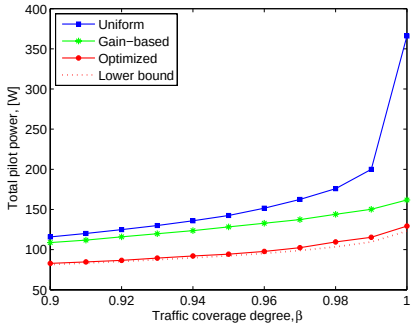


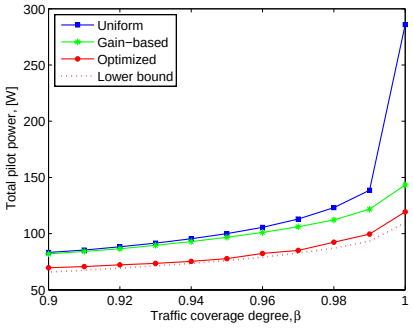
Figure 4.2: Cumulative distribution of cell CPICH power levels in the optimized solutions for Net6 for various traffic coverage degrees.



(a) Test network Net1

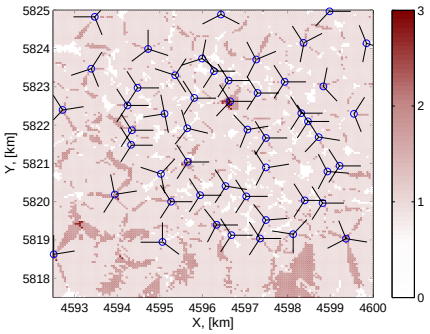


(b) Test network Net6

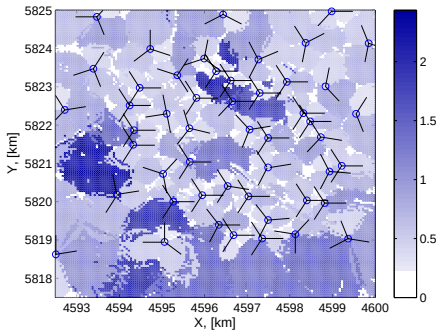


(c) Test network Net7

Figure 4.3: CPICH power consumption versus traffic coverage degree.



(a) Coverage statistics



(b) Best-server CPICH power, (W)

Figure 4.4: Optimized power solution for Net6, $\beta = 0.95$.

Table 4.1: Solutions for various coverage degrees, Net1

Traffic degree β	Uniform		Gain-based		Lagrangian heuristic				
					Integer solution		Lower bound	Gap [%]	CPU [sec]
	Total power [W]	Overlap [%]	Total power [W]	Overlap [%]	Total power, [W]	Overlap [%]			
1.00	71.23	27.64	37.27	8.29	33.53	3.64	33.50	0.11	1.06
0.99	53.27	18.18	33.77	5.24	29.96	2.03	29.83	0.43	2.25
0.98	45.05	12.95	31.68	3.49	28.44	1.67	28.42	0.07	2.16
0.97	41.44	11.05	30.32	2.76	27.37	1.55	27.36	0.03	1.02
0.96	39.51	9.16	28.89	1.96	26.51	1.45	26.46	0.20	2.19
0.95	37.03	7.64	28.28	1.38	25.68	1.16	25.65	0.12	2.12
0.94	34.95	5.45	27.57	0.95	24.94	1.24	24.90	0.14	1.05
0.93	32.41	3.41	26.72	0.44	24.37	0.87	24.27	0.41	1.24
0.92	30.50	1.67	26.41	0.22	23.79	0.58	23.67	0.50	1.12
0.91	29.68	1.16	25.81	0.22	23.21	0.51	23.09	0.51	1.19
0.90	28.85	1.02	25.14	0.15	22.72	0.51	22.54	0.81	0.95

Table 4.2: Solutions for various coverage degrees, Net6

Traffic degree β	Uniform		Gain-based		Lagrangian heuristic				
					Integer solution		Lower bound	Gap [%]	CPU [sec]
	Total power [W]	Overlap [%]	Total power [W]	Overlap [%]	Total power, [W]	Overlap [%]			
1.00	365.95	72.51	161.82	41.01	129.34	26.02	123.03	5.12	119.57
0.99	199.96	46.39	150.22	33.90	115.34	19.46	109.56	5.28	219.61
0.98	175.94	38.81	144.01	30.07	109.56	16.94	103.40	5.96	198.48
0.97	162.48	33.98	137.44	26.70	102.36	14.24	98.92	3.47	180.89
0.96	151.53	29.48	132.85	24.09	97.80	12.15	95.27	2.67	166.39
0.95	142.51	26.01	128.26	21.64	94.33	10.17	92.31	2.19	164.00
0.94	135.86	23.06	123.61	19.26	92.06	9.63	89.76	2.57	158.82
0.93	130.04	20.43	119.82	17.36	89.51	8.49	87.42	2.39	155.25
0.92	124.85	18.38	115.89	15.76	86.62	7.50	85.27	1.58	152.54
0.91	120.14	16.46	111.82	13.95	84.66	7.07	83.25	1.70	145.23
0.90	115.82	14.52	108.74	12.47	82.82	6.61	81.37	1.79	146.13

Table 4.3: Solutions for various coverage degrees, Net7

Traffic degree β	Uniform		Gain-based		Lagrangian heuristic				
					Integer solution		Lower bound	Gap [%]	CPU [sec]
	Total power [W]	Overlap [%]	Total power [W]	Overlap [%]	Total power, [W]	Overlap [%]			
1.00	285.91	65.11	143.32	37.38	119.39	28.72	109.76	8.76	734.96
0.99	138.63	31.87	121.76	27.25	99.59	21.17	93.24	6.81	668.76
0.98	123.12	25.55	112.21	22.30	92.40	17.76	87.14	6.03	632.62
0.97	113.01	21.22	106.07	19.11	85.09	13.79	82.70	2.89	500.44
0.96	105.64	17.78	101.10	16.36	82.39	13.14	79.08	4.18	463.69
0.95	100.01	15.16	96.74	14.33	77.79	10.62	76.07	2.26	437.73
0.94	95.44	13.23	92.93	12.46	75.38	9.78	73.51	2.54	415.72
0.93	91.60	11.36	89.62	10.79	73.57	9.46	71.28	3.20	411.89
0.92	88.36	9.82	86.57	9.33	72.28	9.49	69.30	4.29	405.66
0.91	85.38	8.57	84.19	8.19	70.74	9.13	67.52	4.77	395.58
0.90	83.33	7.65	81.96	7.31	69.68	8.73	65.89	5.75	394.02

becomes relatively small for $\beta = 0.90$. An explanation is that the size of areas that require high CPICH power for coverage is relatively small in these two networks. Having low traffic demand in Net6 and Net7, these areas tend to be not covered in solutions with partial coverage

when $\beta = 0.9$, whilst the required beset-server CPICH power in the rest of the network does not vary a lot. The situation is slightly different in Net1 which is reflected in the results presented in Figure 4.3(a).

Another interesting observation is that the difference between the gain-based power solutions and the optimized CPICH power solutions decreases very slowly when β is lowered in all three networks. This is because both approaches choose approximately the same CPICH coverage areas for the same coverage degree. This observation could be effectively utilized considering that the computational time for the gain-based approach is much shorter than that for finding a near-optimal solution. For example, having an optimal or near-optimal solution for $\beta = \beta_1$ and gain-based solutions for $\beta = \beta_1$ and $\beta = \beta_2$, one could roughly estimate the average CPICH power per cell in an optimal solution for $\beta = \beta_2$. Another possible use of the gain-based approach is to apply it for defining the subset of bins that together satisfy the coverage requirement. A pilot optimization problem is then defined for this subset of bins and solved, for example, by the Lagrangian heuristic that does not involve solving the knapsack problem.

Figures 4.4 show the optimized solution for $\beta = 0.95$ in the city of Berlin (Net6). The solution has been obtained by the Lagrangian heuristic described in Section 4.5. The white pixels in Figure 4.4 are the areas where a mobile terminal cannot detect any CPICH signal at the required RSCP or E_c/I_0 levels. Figure 4.4(a) presents the CPICH coverage map where the color of a bin corresponds to the number of CPICH signals covering this bin. Figure 4.4(b) presents the best-server CPICH power distribution over the area. Figure 4.4(b) indicates also the cell boundaries. Comparing Figure 3.7(b), Figure 4.4(b), and traffic demand distribution shown in Figure A.2(b) (see Appendix A), observe that in the optimized power solution coverage holes follow the traffic distribution and tend to occur mainly in bins with low traffic demand and relatively high required pilot power (due to high interference).

4.7 Conclusions

Several conclusions can be drawn from the computational study presented in this chapter. First, following the Pareto's principle, a slight decrease in the degree of coverage enables considerable reductions in the total amount of pilot power. Coverage versus CPICH power consumption is thus an important trade-off in UMTS network design. Furthermore, for any traffic coverage degree, the optimized CPICH power considerably outperforms the uniform pilot power approach, a strategy commonly adopted in practice by network operators. The power reduction is more significant when β is close to 1.0. This is an important observation because coverage degree values between 0.95 and 1.0 are most relevant from the practical point of view. Another interesting observation is that uniform and gain-based solutions tend to converge when the coverage degree decreases, but there is still a significant gap between the two solutions and the optimized solution.

Chapter 5

Optimization of Radio Base Station Antenna Configuration and Pilot Power

5.1 System Model

The system model in this chapter is very similar to that for the basic pilot power problem presented in Section 3.1. The following changes have been applied in this chapter. Assuming a single directional antenna in each cell, we define set $\mathcal{K}_i$ to represent all possible antenna configurations of mechanical tilt, electrical tilt, and azimuth in cell i, $i \in \mathcal{I}$ (note that in Chapters 3 and 4, CPICH power was viewed as the only changing cell configuration parameter and therefore, set $\mathcal{K}_i$ was used to denote indices of CPICH power levels). A network-wise configuration is denoted by a vector $\mathrm{k} = (k_1, k_2, \ldots, k_{|\mathcal{I}|})$, where $k_i \in \mathcal{K}_i, i \in \mathcal{I}$, i.e., the set of all possible network configurations is defined by the Cartesian product $\mathcal{K}_1 \times \mathcal{K}_2 \times \ldots \times \mathcal{K}_{|\mathcal{I}|}$, and the size of the set is therefore $|\mathcal{K}_1| \cdot |\mathcal{K}_2| \cdot \ldots \cdot |\mathcal{K}_{|\mathcal{I}|}|$ which is an enormous number for a real network.

Unlike in Section 3.1 where the antenna configurations are fixed, in this chapter, the power gain between bin j and the antenna in cell i is not a single value but is given by set $\{g_{ij}^k, k \in \mathcal{K}_i\}$. Given network configuration k, the minimum CPICH power in cell i necessary to cover bin j is therefore a function of the network configuration k and is defined as follows,

$$P_{ij}(\mathrm{k}) = \max\left\{\frac{\gamma_1}{g_{ij}^{k_i}}, \frac{\gamma_0}{g_{ij}^{k_i}} \cdot \left(\sum_{l\in\mathcal{I}} g_{lj}^{k_l}\eta_l^{DL}P_l^{max} + \nu_j + I_j^A\right)\right\} , \tag{5.1}$$

where all parameters, except power gain values, are as defined in Section 3.1. The uniform CPICH power is now also a function of network configuration k, i.e.,

$$P^U(\mathrm{k}) = \max_{j\in\mathcal{J}} \min_{i\in\mathcal{I}} P_{ij}(\mathrm{k}) \ . \tag{5.2}$$

As a result, even finding an optimal solution of uniform CPICH power subject to full coverage is not any more as trivial as in Section 3.4.1.

The goal is to optimize network performance under heavy traffic load consuming all power available at RBSs by finding a good network configuration and defining the CPICH power in each cell subject to full coverage. The optimization problem we tackle is defined in the next section.

5.2 Optimization Problem

Having antenna configuration and CPICH power as decision variables, we formulate the optimization problem as follows.

Objective

- Find network configuration $\mathrm{k}^* = (k_1^*, k_2^*, \ldots, k_{|\mathcal{I}|}^*)$ and a vector $P^{CPICH} = \{P_i^{CPICH}, i \in \mathcal{I}\}$ that minimize the total amount of pilot power in the network, i.e., $\sum_{i \in \mathcal{I}} P_i^{CPICH} \rightarrow \min$.

Constraints

- Full coverage is to be achieved, i.e., for each bin $j \in \mathcal{J}$, there must exist at least one cell i satisfying $P_i^{CPICH} \geq P_{ij}(\mathrm{k}^*)$.
- The pilot power in any cell i is within a given interval, i.e., $\Pi_i^{min} \leq P_i^{CPICH} \leq \Pi_i^{max}$.

We denote this optimization problem by M3. Similar to problems M1 and M2, CPICH power has been chosen for the objective function also in M3, because it influences power consumption of common channels, cell size and load, and reflects the total interference in the network which in turn depends on antenna configurations. This optimization problem is very complex and extremely difficult to solve. In practice, due to the huge search space, the problem is even more difficult than M1 or M2, although all three optimization problems are $\mathcal{NP}$-hard (the $\mathcal{NP}$-hardness property of M3 will be shown later in the section). Therefore, the problem is tackled in two steps. In the first step, the CPICH power is restricted to be uniform in all cells. Thus, in this step the problem reduces to finding a configuration vector k^* minimizing the uniform CPICH power $P^U(\mathrm{k})$. Let us denote the problem by M3u. One particular reason for treating uniform CPICH power, in addition to problem decomposition, is that it is a common practice in currently deployed UMTS networks. The output of the first step is a configuration vector. In the second step, the problem of non-uniform CPICH power is solved under the network configuration vector obtained in the first step.

Finding the optimal configuration k^* for uniform CPICH power is a non-convex optimization problem with numerous local optima and attraction areas varying greatly in size . Moreover, we prove that problem M3u is $\mathcal{NP}$-hard as formalized in Theorem 5.1.

Theorem 5.1. *M3u is $\mathcal{NP}$-hard.*

Proof. An optimization problem is $\mathcal{NP}$-hard if it has an $\mathcal{NP}$-complete recognition version, i.e., a corresponding recognition problem M3u-r is in the class $\mathcal{NP}$, and all other problems in $\mathcal{NP}$ polynomially transform to M3u-r.

> [M3u-r]: Does there exist a network configuration such that the minimum uniform CPICH power providing full CPICH coverage in the network does not exceed P^*?

In other words, given a set of variables $\{v_i^k, i \in \mathcal{I}, k \in \mathcal{K}_i\}$ in which v_i^k is one if configuration k is applied in cell i and zero otherwise, we are interested in finding a setting such that

$$\sum_{k \in \mathcal{K}_i} v_i^k = 1, \qquad i \in \mathcal{I}, \tag{5.3}$$

and for every bin j and its best server i the following inequality holds,

$$\frac{\sum_{k \in \mathcal{K}_i} P^* g_{ij}^k v_i^k}{\sum_{l \in \mathcal{I}} \sum_{k \in \mathcal{K}_l} P^{Tot} g_{lj}^k v_l^k + \nu_j} \geq \gamma_0 \ . \tag{5.4}$$

Obviously, M3u-r is in $\mathcal{NP}$ since every yes instance of M3u-r can be verified in polynomial time $\mathcal{O}(|\mathcal{I}| \times |\mathcal{J}|)$. To prove that all other problems in $\mathcal{NP}$ polynomially transform to M3u-r, it is sufficient to show that a known $\mathcal{NP}$-complete problem polynomially transforms to

M3u-r. We show that a recognition version of a multiple-choice knapsack problem (which is known to be $\mathcal{NP}$-complete) polynomially transforms to M3u-r, i.e., for every instance I_1 of the former we can construct in polynomial time an instance I_2 of M3u-r such that I_1 is a yes instance of the knapsack recognition problem if an only if I_2 is a yes instance of M3u-r.

Consider a recognition version of a multi-dimensional multiple-choice knapsack problem with $\sum_{i\in\mathcal{I}}|\mathcal{K}_i|$ binary variables, $|\mathcal{J}'|$ resource constraints, and $|\mathcal{I}|$ mutually disjoint multiple-choice constraints, where $|\mathcal{I}|$ is the number of groups of items with $|\mathcal{K}_i|$ items in group i, and $\mathcal{J}'$ is the set of resources. Sets $\mathcal{I}$ and $\mathcal{K}_i, i \in \mathcal{I}$ in the recognition version of the considered knapsack problem correspond to those in M3u-r (this will be utilized later for constructing an instance of M3u-r). The entire set of items is given by $\bigcup_{i\in\mathcal{I}}\mathcal{K}_i$ where $\mathcal{K}_{i_1} \cap \mathcal{K}_{i_2} = \emptyset, \forall(i_1, i_2) \in \mathcal{I}\times\mathcal{I}, i_1 \neq i_2$. Mathematically, the recognition version of the multi-dimensional multiple-choice knapsack problem can be formulated as follows,

$$\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} u_i^k v_i^k \geq u^* \tag{5.5a}$$

$$\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i} w_{ij}^k v_i^k \leq w_j^* \qquad j \in \mathcal{J}' \tag{5.5b}$$

$$\sum_{k\in\mathcal{K}_i} v_i^k = 1 \qquad i \in \mathcal{I} \tag{5.5c}$$

$$v_i^k \in \{0,1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{5.5d}$$

where u^* is a given total value of the selection, u_i^k is the value of item k in group i, w_j^* is a given limit on resource j, and w_{ij}^k is the amount of resource j consumed by item k in group i. All the parameters (u_i^k, u^*, w_{ij}^k, w_j^*) are non-negative numbers.

The corresponding instance of M3u-r has a set of bins $\mathcal{J}$ where $|\mathcal{J}| = |\mathcal{J}'| + 1$, a set of cells $\mathcal{I}$, and a set of possible configurations $\mathcal{K}_i$ in cell i, $i \in \mathcal{I}$. Consider a network where the best serving cell is known for each bin in advance, although we do not put any restriction on how this cell is to be chosen, i.e., it can be any cell from set $\mathcal{I}$. (Later we will show that it is possible to ensure in the constructed instance that the selected cell is the best server.) Then, the mathematical formulation of M3u-r is as follows.

$$\frac{\sum_{k\in\mathcal{K}_i} P^* g_{ij}^k v_i^k}{\sum_{l\in\mathcal{I}}\sum_{k\in\mathcal{K}_l} P^{Tot} g_{lj}^k v_l^k + \nu_j} \geq \gamma_0 \qquad j \in \mathcal{J}, i \in \mathcal{I}_j \tag{5.6a}$$

$$\sum_{k\in\mathcal{K}_i} v_i^k = 1 \qquad i \in \mathcal{I} \tag{5.6b}$$

$$v_i^k \in \{0,1\} \qquad i \in \mathcal{I}, k \in \mathcal{K}_i \tag{5.6c}$$

where $\mathcal{I}_j$ is the set of a single element which is the best serving cell in bin j ($|\mathcal{I}_j| = 1$).

Consider the following transformation of constraints (5.6a):

$$\begin{aligned}
&\sum_{k\in\mathcal{K}_i} P^* g_{ij}^k v_i^k \geq \gamma_0 \cdot \left(\sum_{l\in\mathcal{I}}\sum_{k\in\mathcal{K}_l} P^{Tot} g_{lj}^k v_l^k + \nu_j\right) \\
&\iff \sum_{l\in\mathcal{I}}\sum_{k\in\mathcal{K}_l} g_{lj}^k v_l^k - \frac{P^*}{\gamma_0 P^{Tot}} \cdot \sum_{k\in\mathcal{K}_i} g_{ij}^k v_i^k \leq -\frac{\nu_j}{P^{Tot}} \\
&\iff \sum_{\substack{l\in\mathcal{I}:\\ l\neq i}}\sum_{k\in\mathcal{K}_l} g_{lj}^k v_l^k + \left(1 - \frac{P^*}{\gamma_0 P^{Tot}}\right) \cdot \sum_{k\in\mathcal{K}_i} g_{ij}^k v_i^k \leq -\frac{\nu_j}{P^{Tot}}
\end{aligned} \tag{5.7}$$

Note that since the right-hand side of (5.7) is negative, the coefficient in the round brackets in the left-hand side must be negative, i.e., $\frac{P^*}{P^{Tot}} \geq \gamma_0$ has to be satisfied by the instance.

Further, we obtain an equivalent set of constraints with non-negative coefficients. For each bin j, we introduce a (positive) parameter b_j and derive the following transformation of (5.7).

$$\sum_{\substack{l\in\mathcal{I}:\\ l\neq i}}\sum_{k\in\mathcal{K}_l} g_{lj}^k v_l^k + \left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\sum_{k\in\mathcal{K}_i}(-g_{ij}^k)v_i^k + \left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\cdot b_j \leq \left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\cdot b_j - \frac{\nu_j}{P^{Tot}} \tag{5.8}$$

Next, utilizing constraints (5.6b), we substitute $\left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\cdot b_j$ with $\left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\cdot b_j \cdot \sum_{k\in\mathcal{K}_i} v_i^k$ and obtain the following inequality,

$$\sum_{\substack{l\in\mathcal{I}:\\ l\neq i}}\sum_{k\in\mathcal{K}_l} g_{lj}^k v_l^k + \left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right)\sum_{k\in\mathcal{K}_i}(b_j - g_{ij}^k)v_i^k \leq \frac{P^*\cdot b_j}{\gamma_0 P^{Tot}} - b_j - \frac{\nu_j}{P^{Tot}} \tag{5.9}$$

Note that (5.9) is valid only if (5.6b) and (5.6a) are valid. Observe that if $b_j \geq \max_{k\in\mathcal{K}_i} g_{ij}^k$, all coefficients in the left-hand side are non-negative (provided that $\frac{P^*}{\gamma_0 P^{Tot}} > 1$). Because $0 \leq g_{ij}^k \leq 1, i \in \mathcal{I}, j \in \mathcal{J}, k \in \mathcal{K}_i$, we choose $b_j = 1$ for all $j \in \mathcal{J}$.

Consider the first element of $\mathcal{J}$ and denote it by j_1. For bin j_1, we construct the CIR constraint from an inequality derived by scaling (5.5a) and subtracting it from the sum of all constraints (5.5c) of the knapsack recognition problem:

$$\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_i}(1-\bar{u}_i^k)v_i^k \leq |\mathcal{I}| - \bar{u}^* \ , \tag{5.10}$$

where $\bar{u}_i^k = \dfrac{u_i^k}{\max\limits_{i\in\mathcal{I}}\max\limits_{k\in\mathcal{K}_i} u_i^k} \leq 1$ and $\bar{u}^* = \dfrac{u^*}{\max\limits_{i\in\mathcal{I}}\max\limits_{k\in\mathcal{K}_i} u_i^k}$.

Power gain values for bin j_1 with respect to each cell are chosen as follows,

- $g_{lj_1}^k = 1 - \bar{u}_l^k$, for all $l \in \mathcal{I}\setminus\mathcal{I}_{j_1}, k \in \mathcal{K}_l$,
- $g_{ij_1}^k = 1 - \frac{1-\bar{u}_i^k}{\frac{P^*}{\gamma_0 P^{Tot}}-1}$, for all $i \in \mathcal{I}_{j_1}, k \in \mathcal{K}_i$.

Note that $g_{lj_1}^k < 1,\ l \in \mathcal{I}\setminus\mathcal{I}_{j_1}, k \in \mathcal{K}_l$, since in the knapsack problem it is always possible to ensure (in preprocessing) that $\bar{u}_l^k > 0$ holds for all indexes l and k. Therefore, by choosing properly parameters P^*, P^{Tot}, and γ_0 such that

$$\min_{k\in\mathcal{K}_i}\left\{1 - \frac{1-\bar{u}_i^k}{\frac{P^*}{\gamma_0 P^{Tot}}-1}\right\} \geq \max_{l\in\mathcal{I}\setminus\mathcal{I}_{j_1}}\max_{k\in\mathcal{K}_l}\{1-\bar{u}_l^k\} \ , \tag{5.11}$$

we can always ensure that cell $i \in \mathcal{I}_{j_1}$ is the best server for bin j_1.

For bins $j \in \mathcal{J}\setminus\{j_1\}$ we define $|\mathcal{J}'|$ corresponding constraints in the knapsack recognition problem (constraints (5.5b)). The corresponding power gain values of M3u-r for each bin $j \in \mathcal{J}\setminus\{j_1\}$ are derived below:

- $g_{lj}^k = \bar{w}_{lj}^k$, for all $j \in \mathcal{J}\setminus\{j_1\}, l \in \mathcal{I}\setminus\mathcal{I}_j, k \in \mathcal{K}_l$,
- $g_{ij}^k = 1 - \frac{\bar{w}_{ij}^k}{\frac{P^*}{\gamma_0 P^{Tot}}-1}$, for all $j \in \mathcal{J}\setminus\{j_1\}, i \in \mathcal{I}_j, k \in \mathcal{K}_i$,

where $\bar{w}_{ij}^k = \dfrac{w_{ij}^k}{w_j^*},\ i \in \mathcal{I}, j \in \mathcal{J} \setminus \{j_1\}$. Note that elements j for which $w_{ij}^k = w_j^*$ are typically processed separately in the preprocessing step and are not included in the problem to be solved. Therefore, $g_{lj}^k < 1$ for all $j \in \mathcal{J} \setminus \{j_1\}, l \in \mathcal{I} \setminus \mathcal{I}_j, k \in \mathcal{K}_l$. For any bin $j \in \mathcal{J} \setminus \{j_1\}$, cell $i \in \mathcal{I}_j$ is thus the best server when

$$\min_{k \in \mathcal{K}_i} \left\{ 1 - \frac{\bar{w}_{ij}^k}{\frac{P^*}{\gamma_0 P^{Tot}} - 1} \right\} \geq \max_{l \in \mathcal{I} \setminus \mathcal{I}_j} \max_{k \in \mathcal{K}_l} \bar{w}_{lj}^k \ . \tag{5.12}$$

From (5.11) and (5.12), we obtain a single constraint for parameters P^*, P^{Tot}, and γ_0 as follows:

$$\frac{P^*}{\gamma_0 P^{Tot}} \geq 1 + \max \left\{ \frac{1 - \min\limits_{k \in \mathcal{K}_{\mathcal{I}_{j_1}}} \bar{u}_{\mathcal{I}_{j_1}}^k}{1 - \max\limits_{l \in \mathcal{I} \setminus \mathcal{I}_{j_1}} \max\limits_{k \in \mathcal{K}_l} \{1 - \bar{u}_l^k\}}, \max_{j \in \mathcal{J} \setminus \{j_1\}} \frac{\max\limits_{k \in \mathcal{K}_{\mathcal{I}_j}} \bar{w}_{\mathcal{I}_j j}^k}{1 - \max\limits_{l \in \mathcal{I} \setminus \mathcal{I}_j} \max\limits_{k \in \mathcal{K}_l} \bar{w}_{lj}^k} \right\} \ . \tag{5.13}$$

Observe that if inequality (5.13) holds, requirement $\frac{P^*}{P^{Tot}} \geq \gamma_0$ is redundant.

Once parameters P^*, P^{Tot}, and γ_0 are set, the thermal noise parameters can be found from the following two sets of equations,

- $\left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right) - \frac{\nu_{j_1}}{P^{Tot}} = |\mathcal{I}| - \bar{u}^* \ ,$
- $\left(\frac{P^*}{\gamma_0 P^{Tot}} - 1\right) - \frac{\nu_j}{P^{Tot}} = 1, \quad \text{for all } j \in \mathcal{J} \setminus \{j_1\} \ .$

The described transformation can be clearly done in polynomial time since parameters P^*, γ_0, and P^{Tot} are chosen to satisfy inequality (5.13), and the noise parameters are obtained by solving simple equations (one equation for each $\nu_j, j \in \mathcal{J}$). Moreover, constraints (5.6b) in M3u-r and constraints (5.5c) in the knapsack recognition problem are identical. By construction, it is true that the knapsack recognition problem is a yes instance if and only if the constructed instance of M3u-r is. Thus, we have showed that the knapsack recognition problem polynomially transforms to M3u-r. Hence the conclusion. □

Corollary 5.1. *M3 is $\mathcal{NP}$-hard.*

Proof. Given the results of Theorem 5.1, the $\mathcal{NP}$-hardness of M3 is straightforward since M3u is a special case of M3. □

5.3 Solution Approach

As mentioned in Section 5.2, the solution approach for optimization problem M3 consists of two steps:

Step 1. Solve M3u, i.e., find configuration k* minimizing P^U(k);

Step 2. For fixed configuration k* obtained in Step 1, solve M1 to find non-uniform CPICH power.

In the first step, the problem of minimizing uniform CPICH power is solved using simulated annealing, a probabilistic meta-heuristic algorithm for searching in the solution space of hard optimization problems (see Section 1.2.3 for a general description of the simulated annealing algorithm and references). In the second step, the optimization algorithm presented in Section 3.6 is applied to optimize non-uniform CPICH power under given network configuration vector.

As mentioned in Section 1.2.3, the simulated annealing algorithm operates using the neighborhood principle and occasionally accepts non-improving solutions in order to escape from local optima. The neighborhood definition and the algorithm control parameters that define the probability of accepting worse solutions, however, must be adapted to each specific problem since they greatly affect algorithm performance. Sections 5.3.1 and 5.3.2 address the two issues.

5.3.1 Generating New Solutions

The algorithm starts from a given initial network configuration k^0 and iteratively moves in the search space. In each iteration, the decision about the next move is made after generating and testing a new solution (network configuration). Thus, generating new solutions is a crucial part of the algorithm. To create a new configuration, the algorithm changes the antenna configuration of one cell, say cell i', in the current network configuration k. The new network configuration is thus $\mathrm{k}' = (k_1, k_2, \ldots, k'_{i'}, \ldots, k_{|\mathcal{I}|})$, where $k'_{i'}$ is the new configuration in cell i'.

Cell i' is chosen randomly with an exponentially distributed probability from a *candidate set*. The candidate set contains the μ first cells when cells are put in a list sorted in descending order by their power gains with respect to the *bottleneck bin* b of the current network configuration k. A bottleneck bin is the bin defining the minimum uniform CPICH power satisfying the full coverage requirement. Typically, cells with large power gain values provide CPICH coverage in bin b with less CPICH power and therefore, have a significant influence on the bin coverage. Parameter μ is defined as a function (implemented as **findCandSetSize**(n_{bad}) in Algorithm I.5) of the number n_{bad} of unsuccessful moves of the simulated annealing algorithm, such that less unsuccessful iterations results in smaller μ. In the set of candidate cells, cell i' is determined by random index **round**(μ^{1-r}), where r is a random number in the range $[0, 1]$, and **round**$(\cdot)$ is a function that rounds the input parameter to the nearest integer.

A new configuration $k'_{i'}$ in cell i' is chosen randomly from set $\mathcal{K}'_{i'}$. One or several trials may be performed, and in each trial, the current configuration of cell i' and those previously tested configurations are excluded. To speed up the algorithm, each chosen configuration is evaluated only with respect to the minimum CPICH power for the bottleneck bin b, i.e., $P_{i'b}(\mathrm{k})$, without computing the minimum uniform CPICH power for the entire area for network configuration k'. This rough evaluation is based on the observation that a configuration that increases the minimum CPICH power in bin b, increases also the uniform CPICH power in the new network configuration. A configuration that improves the CPICH power in bin b is accepted directly as the new network configuration. Otherwise, another attempt is performed, if the maximum allowed number of attempts is not exceeded. The algorithm for selecting a new network configuration is outlined in Algorithm I.5 and requires the following input parameters,

- → k is the current network configuration,
- → n_{bad} is the current number of non-improving algorithm moves,
- → $maxTrials$ is the maximum number of trial configurations for the selected cell i' ($maxTrials \leq |\mathcal{K}_{i'}|$).

The algorithm calls an auxiliary function, **rand**(), which is a function that returns a random number distributed uniformly in $[0, 1]$.

5.3.2 Algorithm Parameters

In each iteration, the algorithm always accepts a new network configuration as its next move if this does not increase the uniform pilot power. If the new configuration is worse than

Algorithm I.5 Generating a new network configuration

Input: $\mathrm{k}, n_{bad}, maxTrials$
Output: k'

$\mathrm{k}' \Leftarrow \mathrm{k}$
$[b, c] \Leftarrow \arg\max_{j \in \mathcal{J}} \min_{i \in \mathcal{I}} P_{ij}(\mathrm{k})$ // Find bottleneck bin b and covering cell c
$\bar{\mathcal{I}} \Leftarrow \mathbf{sortD}(\{g_{ib}^{k_i}, i \in \mathcal{I}\})$ // Sort cells in descending order by their gains in bin b
$\mu \Leftarrow \mathbf{findCandSetSize}(n_{bad})$
$ind \Leftarrow \mathbf{round}(\mu^{1-\mathbf{rand}()})$
$i' \Leftarrow \mathbf{get}(ind, \bar{\mathcal{I}})$ // Get a candidate cell (a cell with index ind in the sorted list)
$\bar{\mathcal{K}}_{i'} \Leftarrow \mathcal{K}_{i'} \setminus \{k_{i'}\}$
$trials \Leftarrow 0$
$gMin \Leftarrow g_{cb}^{k_c}$
$gMax \Leftarrow 0$
while $trials < maxTrials$ **do**
    $ind \Leftarrow \mathbf{round}((|\bar{\mathcal{K}}_{i'}| - 1) \cdot \mathbf{rand}()) + 1$ // Find a candidate configuration in cell i'
    $\kappa \Leftarrow \mathbf{get}(ind, \bar{\mathcal{K}}_{i'})$
    if $i' == c$ **then**
        if $g_{i'b}^{\kappa} \geq g_{cb}^{k_c}$ **then**
            $k'_{i'} \Leftarrow \kappa$
            break // Break the while loop
        else
            if $g_{i'b}^{\kappa} < gMin$ **then**
                $k'_{i'} \Leftarrow \kappa$
                $gMin \Leftarrow g_{i'b}^{\kappa}$
            end if
        end if
    else
        if $g_{i'b}^{\kappa} \leq g_{cb}^{k_c}$ **then**
            $k'_{i'} \Leftarrow \kappa$
            break // Break the while loop
        else
            if $g_{i'b}^{\kappa} > gMax$ **then**
                $k'_{i'} \Leftarrow \kappa$
                $gMax \Leftarrow g_{i'b}^{\kappa}$
            end if
        end if
    end if
    $trial \Leftarrow trials + 1$
    $\bar{\mathcal{K}}_{i'} \Leftarrow \bar{\mathcal{K}}_{i'} \setminus \{\kappa\}$
end while

the current one, it is accepted with probability p. The probability is determined by two factors. The first is the difference δ in objective function values (uniform pilot power). A non-improving solution with small (positive) δ has a higher probability of being accepted than a solution with larger δ. The second factor is the temperature parameter T. Higher T means higher probability. Algorithmically, the probability p is calculated as

$$p = e^{-\frac{\delta}{T}}. \tag{5.14}$$

The temperature parameter T is decreased gradually. Eventually, the process converges to a frozen state in which the probability of accepting inferior solutions is almost zero. In a practical implementation, simulated annealing terminates if a maximum allowed number

of iterations has been reached, or if no improving move has been found in a number of consecutive iterations.

In our implementation, δ is the relative difference in uniform CPICH power, that is,

$$\delta = \frac{P^U(\mathrm{k}') - P^U(\mathrm{k})}{P^U(\mathrm{k})} \ , \tag{5.15}$$

where k and k$'$ are the current and the new configuration, respectively. Using the relative difference of the objective function when computing the probability p instead of the absolute deviation as in the original simulated annealing algorithm makes the algorithm less sensitive to variation in problem instances. In particular, the choice of initial temperature T_0 and temperature reduction factor f, the two parameters of crucial importance for the algorithm performance, is determined by the desired maximum number of algorithm iterations, the initial and the last-iteration values of probability p and relative difference δ, but does not depend on the range of the objective function values.

Let N denote the maximum allowed number of iterations. In addition to N, algorithm parameter specification consists of two tuples: (p_0, δ_0) and (p_N, δ_N). Here, p_0 is the probability of accepting a solution with relative difference δ_0 in the first iteration, and p_N and δ_N are the corresponding entities after N iterations. With (p_0, δ_0) and (p_N, δ_N), we derive two temperature values from (5.14):

$$T_0 = -\frac{\delta_0}{\ln p_0} \ , \tag{5.16}$$

$$T_N = -\frac{\delta_N}{\ln p_N} \ . \tag{5.17}$$

The algorithm uses T_0 as the initial temperature, which is then reduced by a scaling factor f $(0 < f < 1)$ in every iteration, such that the temperature becomes T_N after N iterations. Given T_0 and T_N, the temperature reduction factor can be found as follows,

$$f = \left(\frac{T_N}{T_0}\right)^{\frac{1}{N}} \ . \tag{5.18}$$

5.3.3 The Optimization Algorithm

The implementation of the simulated annealing algorithm is outlined in Algorithm I.6. The input parameters of the algorithm are listed and commented below.

- → k^0 is the starting network configuration,
- → T_0 is the initial temperature,
- → f is the temperature reduction factor,
- → N is the maximum number of iterations,
- → N_{bad} is the maximum number of consecutive non-improving iterations.

The output of the algorithm are the best found configuration k^* and the corresponding uniform CPICH power $P^* = P^U(\mathrm{k}^*)$. Function **findOptUnif**$(\cdot)$ computes the minimum uniform CPICH power for a given network configuration by equation (5.2). Function **rand**() is, as previously, a function that returns a random number drawn from a uniform distribution on the unit interval, and function **generateNewConfig**$(\cdot)$ generates a new network configuration as described in Section 5.3.1.

Algorithm I.6 The simulated annealing algorithm for optimizing uniform pilot power by adjusting antenna configurations

Input: $\mathrm{k}^0, T_0, f, N, N_{bad}, maxTrials$
Output: k^*, P^*

$\mathrm{k} \Leftarrow \mathrm{k}^0$
$T \Leftarrow T_0$ // Current temperature
$n \Leftarrow 0$ // Current number of iterations
$n_{bad} \Leftarrow 0$ // Current number of consecutive non-improving iterations
$P \Leftarrow$ **findOptUnif**(k) // Current objective function value
$P^* \Leftarrow P$ // Best objective function value
while $(n < N)\&(n_{bad} < N_{bad})$ **do**
    $\mathrm{k}' \Leftarrow$ **generateNewConfig**$(\mathrm{k}, n_{bad}, maxTrials)$ // Generate a new solution
    $P' \Leftarrow$ **findOptUnif**(k') // Compute new objective function value
    $\delta = \frac{P'-P}{P}$
    if $\delta < 0$ **then**
        $P \Leftarrow P'$ // Accept the new solution
        $\mathrm{k} \Leftarrow \mathrm{k}'$
        if $P < P^*$ **then**
            $P^* \Leftarrow P$ // Update the best solution
            $\mathrm{k}^* \Leftarrow \mathrm{k}$
            $n_{bad} \Leftarrow 0$
        end if
    else
        if rand() $\leq e^{-\frac{\delta}{T}}$ **then**
            $P \Leftarrow P'$ // Accept the new solution
            $\mathrm{k} \Leftarrow \mathrm{k}'$
            $n_{bad} \Leftarrow n_{bad} + 1$
        end if
    end if
    $T \Leftarrow f \cdot T$
    $n \Leftarrow n + 1$
end while

5.4 Performance Metrics

In this section, we present a set of performance metrics we use to evaluate the obtained network configurations.

CPICH Transmit Power

CPICH power is the objective of our optimization problem. A relatively small value of uniform pilot power that can guarantee full CPICH coverage is an indication of a good network configuration solution with well-chosen RBS locations, balanced cell sizes, and well-planned cell isolation resulting in low total DL interference. In addition to CPICH, a cell uses a number of other common channels, and the total amount of power of these channels is typically set in proportion to that of CPICH (see Section 2.3.2). We assume that the total transmit power of the other common channels amounts to 80 % of the CPICH power. Thus, the total power consumed by all DL common channels in cell i is $1.8 \cdot P_i^{CPICH}$. As has been discussed in Section 2.3.2, a reduction in CPICH power may give a significant capacity improvement by leaving more power to user traffic.

CPICH Coverage Overlap

The CPICH coverage overlap is estimated as the ratio between the size of the area with overlapping cells and the total area size. We compute two coverage overlap values. The first is the total relative size of the areas where at least two received CPICH signals satisfy the minimum E_c/I_0 requirement. These areas represent potential soft and/or softer handover areas. The second value is computed for the areas with four or more CPICH signals satisfying the CIR threshold, i.e., areas that potentially may suffer from high soft handover overhead and pilot pollution.

DL Load Factor

To estimate the DL air interface load, we use the *DL load factor*, defined for each cell as the ratio between the total allocated transmit power and the maximum transmit power available in the cell, i.e.,

$$\eta_i^{DL} = \frac{P_i^{Tot}}{P_i^{max}} \ . \tag{5.19}$$

Given a user distribution and traffic information for each service type, and assuming that the serving RBS for each user is known, the total transmit power allocated in cell i can be found as the sum of transmit powers of all DL common channels and dedicated traffic channels needed to support all users served by the cell, that is,

$$P_i^{Tot} = \sum_{j \in \bar{\mathcal{J}}_i} \sum_{s \in \mathcal{S}} d_j^s \upsilon^s p_{ij}^s + P_i^{CPICH} + P_i^{com} \ , \tag{5.20}$$

where $\bar{\mathcal{J}}_i$ is the set of bins for which i is the best server, d_j^s is the number of users in bin j requesting service s, $\mathcal{S}$ is the set of services, υ^s is the activity factor of service s, p_{ij}^s is the minimum transmit power needed in the cell to provide a user in bin j with service s, and P_i^{com} is the total transmit power of all DL common channels of cell i except CPICH (by the aforementioned assumption, $P_i^{com} = 0.8 \cdot P_i^{CPICH}$). In our numerical experiments, we assume that a user is always served by the best server. For a given network configuration, the minimum transmit power needed to support a user in bin j can be found from the following signal-to-noise ratio inequality,

$$\frac{g_{ij} p_{ij}^s}{(1-\alpha_j) g_{ij} (P_i^{Tot} - \upsilon^s p_{ij}^s) + \sum_{l \neq i} g_{lj} P_l^{Tot} + \nu_j} \geq \gamma^s \ , \tag{5.21}$$

where α_j is the orthogonality factor in bin j. The threshold value γ^s is defined as $\gamma^s = R^s/W \cdot (E_b/N_0)_{target}^s$, where $W = 3.84\,\text{Mcps}$ is the WCDMA chip rate, R^s is the bit rate of service s, and $(E_b/N_0)_{target}^s$ is the E_b/N_0 target of service s.

From (5.21), we find the minimum DL power needed to support a user in bin j,

$$p_{ij}^s = \frac{\gamma^s}{g_{ij}} \cdot \frac{(1-\alpha_j) g_{ij} P_i^{Tot} + \sum_{l \neq i} g_{lj} P_l^{Tot} + \nu_j}{1 + (1-\alpha_j) \upsilon^s \gamma^s} \ , \tag{5.22}$$

which after substituting into (5.20) gives us the following equation for the DL transmit power in cell i,

$$P_i^{Tot} = P_i^{CPICH} + P_i^{com} + \sum_{j \in \bar{\mathcal{J}}_i} \sum_{s \in \mathcal{S}} d_j^s \phi_j^s \cdot \frac{(1-\alpha_j) g_{ij} P_i^{Tot} + \sum_{l \neq i} g_{lj} P_l^{Tot} + \nu_j}{g_{ij}} \ , \tag{5.23}$$

where ϕ_j^s is defined by

$$\phi_j^s = \frac{v^s \gamma^s}{1 + (1 - \alpha_j) v^s \gamma^s} \ . \tag{5.24}$$

Thus, we obtain a system of linear equations of type (5.23) with a vector of unknown transmit power levels $P_i^{Tot}, i \in \mathcal{I}$. The obtained system can be solved by a standard linear system solver. However, if it is assumed that all users must be served, the system is not always guaranteed to have a feasible solution due to the constraint on the maximum transmit power of a cell. A situation where some users are refused the service due to cell overloading is not taken into account in this formulation.

We extend the described system of linear equations to an optimization problem in which we find the total transmit power of each cell such that the number of cells with at least one blocked user is minimized (see Appendix B for more details). The resulting optimization problem can quickly be solved using a standard linear integer programming solver, e.g., ILOG CPLEX. Once the total amount of DL transmit power of each cell is known, we can compute the DL load factor of the cell by (5.19). Other approaches for computing the DL transmit power levels and estimating the DL load can be found, for example, in [15, 16, 59].

Highly Loaded and Overloaded Cells

Based on the assumption that in a stable network the DL cell load must be kept below 70 % [31], the set of highly loaded cells are those for which $\eta_i^{DL} > 0.7$. Also, we count the number of cells with excessive demand in the total transmit power, i.e., cells in which the total amount of DL transmit power needed to support the specified amount of user traffic exceeds P_i^{max} resulting in blocking some users. These cells are further referred to as overloaded cells.

5.5 Numerical Experiments

5.5.1 Test Network

Numerical experiments have been performed for a test network originating from a planning scenario for the downtown of Lisbon [31] (see Net8 in Appendix A). Some statistics of the test network and the parameter setting used in our computational experiments are summarized in Tables A.2 and A.3.

For the sake of simplicity, we assume that all antennas in the network are of the same type and for each antenna, its location and height are known. The antenna parameters are given in Table 5.1. Given path loss predictions for isotropic antennas [31] and horizontal and vertical antenna diagrams [35], the attenuation values at every point of the service area for each antenna configuration can be computed as a sum of three values: Path loss for an isotropic antenna, antenna gain, and directional loss. Directional losses have been calculated using linear interpolation of four sample points from the antenna diagrams [13] (see [23, 10, 42, 55] for alternative approaches for predicting 3D antenna radiation patterns).

Figure 5.1(a) shows the best-server isotropic antenna prediction for the network. Figure 5.1(b) demonstrates attenuation prediction for the *reference antenna configurations*, where

Table 5.1: Radio base station antenna characteristics

Antenna characteristic	Value
Antenna type	Kathrein 742265
Frequency	2110 MHz
Antenna gain	18.5 dBi
Polarization	+45°
Half-power beam width	+65°
Adjustable electrical downtilt range	[0°, 6°]

antenna azimuth values are those originally given [31] and no downtilt is present. The color of each pixel in the figure represents the attenuation value between the pixel and the best server, i.e., the RBS having the smallest attenuation at the pixel. The reference antenna configuration is a part of the *reference network configuration* in which the CPICH power is set to 10 % of the maximum RBS DL transmit power.

During the optimization process, the set of possible antenna downtilt degrees is $\{0°, 1°, \ldots, 6°\}$ for both mechanical and electrical tilting. Antenna azimuth can be adjusted within the range $[-10°, 10°]$ with a five-degree step relative to the initial direction, i.e., relative azimuth values are $\{-10°, -5°, 0°, 5°, 10°\}$. When mechanical and electrical tilt are combined, preference is given to the latter, that is, mechanical downtilt is used only if electrical downtilt has reached its six-degree maximum. For example, the total downtilt of 11° means that there have been applied electrical tilt of 6° and mechanical tilt of 5°. This is because electrical tilting, in addition to being less costly than mechanical tilting, offers better performance in a capacity-limited urban environment.

We do not explicitly model either data traffic or user distribution in our optimization model since we aim at providing full coverage of the service area under the worst-case interference assumption. This allows us to obtain solutions with CPICH coverage guarantee for any traffic scenario. However, from a practical point of view, it is desirable to evaluate the obtained solutions for a realistic traffic scenario with non-uniform user distribution in order to study, for example, the effect of optimization on the cell loads. Therefore, we examine optimized network configurations for a traffic load modeled by a snapshot [31]. The total number of non-uniformly distributed active users of all service types in the snapshot is 2618, which corresponds to a highly loaded city traffic scenario.

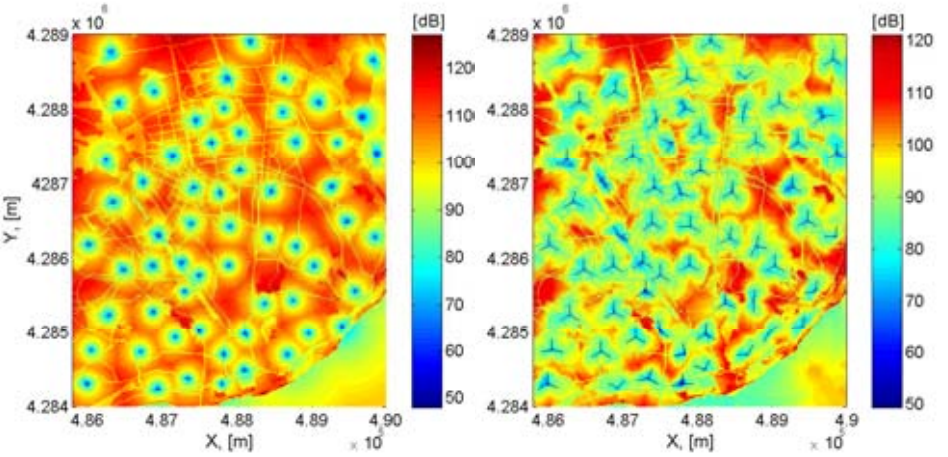
We consider eight services in four groups. The bit rate and the E_b/N_0 target value associated with each of the services are specified below. The DL activity factor is 0.58 for speech, and 1.0 for all data services.

- Conversational
  - Speech telephony (12.2 kbps, 5.5 dB)
  - Video telephony (64 kbps, 4.7 dB)
- Streaming
  - Streaming multimedia (64 kbps, 4.2 dB)
- Interactive
  - Web browsing (32 kbps, 1.9 dB)
  - Location-based service (32 kbps, 3.35 dB)
- Background
  - MMS (32 kbps, 3.35 dB)
  - E-mail (32 kbps, 3.35 dB)
  - File download (64 kbps, 1.9 dB)

In our traffic scenario, the traffic mix is represented by 67.86 % of speech telephony users, 20.38 % of data streaming users, 6.68 % of video telephony users, and 4.75 % of users downloading files. The rest of active users are distributed among the other four services.

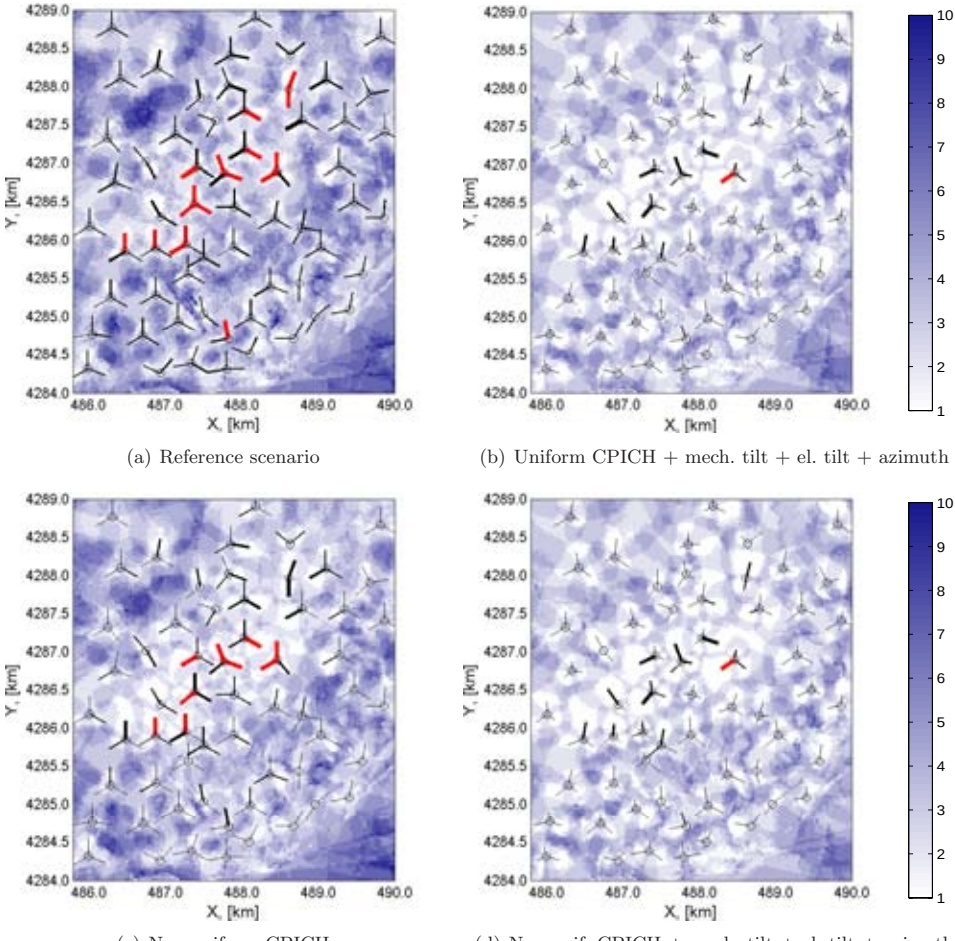
5.5.2 Optimization Results

Results presented in this section have been obtained from numerical experiments with different combinations of optimization parameters. The main goal of the experiments has been to investigate how much improvement in network performance can be achieved by optimizing the CPICH power and antenna configuration. The obtained solutions have been evaluated using performance metrics presented in Section 5.4. The results are summarized in Table 5.2. All computational experiments have been conducted on a moderate-speed computer (Pentium 4-M notebook with a 1.8 GHz CPU and 1 GB RAM).



(a) Isotropic prediction (b) Prediction for the initial network configuration

Figure 5.1: Best-server path loss prediction maps.



(a) Reference scenario (b) Uniform CPICH + mech. tilt + el. tilt + azimuth

(c) Non-uniform CPICH (d) Non-unif. CPICH + mech. tilt + el. tilt + azimuth

Figure 5.2: Coverage and load statistics for selected solutions.

Table 5.2: Results of performance evaluation of different solutions

Optimization scenarios	Common channels' transmit power, [W]		Coverage overlap [%]		Number of highly loaded cells	Number of over-loaded cells	Average DL load factor
	CPICH	Total	≥2 cells	≥4 cells			
Reference network configuration	2.00	3.60	0.93	0.48	24	17	0.41
Uniform CPICH power							
Reference antenna configuration	1.89	3.40	0.93	0.48	24	16	0.40
Azimuth	1.31	2.36	0.91	0.44	19	12	0.32
Mechanical tilt	0.93	1.67	0.89	0.38	8	7	0.20
Electrical tilt	0.82	1.48	0.87	0.34	5	2	0.17
Mechanical tilt + azimuth	0.84	1.51	0.88	0.36	6	4	0.18
Electrical tilt + azimuth	0.74	1.33	0.85	0.31	4	3	0.15
Mech. tilt + el. tilt	0.75	1.35	0.78	0.24	4	4	0.17
Mech. tilt + el. tilt + azimuth	0.71	1.27	0.79	0.26	3	1	0.14
Non-uniform CPICH power							
Reference antenna configuration	0.88	1.58	0.89	0.40	16	9	0.26
Azimuth	0.88	1.58	0.89	0.41	18	11	0.26
Mechanical tilt	0.73	1.32	0.89	0.37	7	6	0.18
Electrical tilt	0.68	1.22	0.86	0.34	5	1	0.15
Mechanical tilt + azimuth	0.71	1.28	0.87	0.36	5	4	0.16
Electrical tilt + azimuth	0.65	1.17	0.85	0.31	4	3	0.13
Mech. tilt + el. tilt	0.64	1.16	0.78	0.23	4	3	0.16
Mech. tilt + el. tilt + azimuth	0.62	1.12	0.78	0.24	3	1	0.13

Following the problem decomposition approach in Section 5.3, performance evaluation consists of two parts. The first part contains results obtained by applying the simulated annealing algorithm to minimize uniform CPICH power. To study the effect of mechanical tilt, electrical tilt, and antenna azimuth, the algorithm has been run considering each of the configuration parameters, as well as combinations of them. The second part of performance evaluation addresses solutions of non-uniform CPICH power, obtained by Algorithm I.1 presented in Section 3.6 for the solutions in part one. In each of the two parts, we also evaluate the solution of optimizing CPICH power for the reference antenna configuration. Note that optimizing uniform CPICH power for the reference antenna configuration simply amounts to

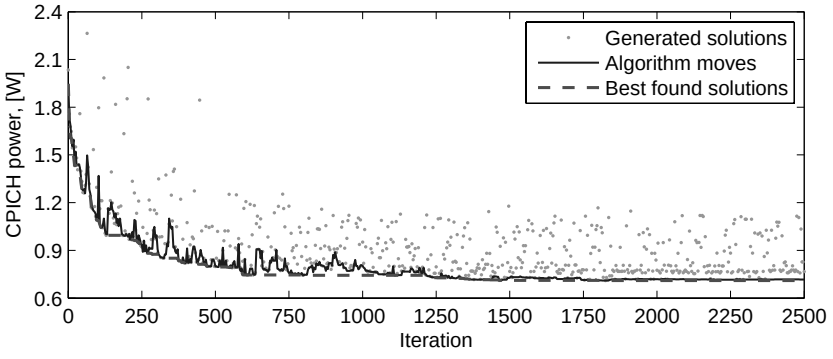


Figure 5.3: Convergence plot of the simulated annealing algorithm when optimizing uniform CPICH power, antenna mechanical and electrical tilts, and antenna azimuth.

computing the CPICH power by equation (5.2). Beside solutions of uniform and non-uniform CPICH power, Table 5.2 presents also performance metrics for the reference network configuration which uses the reference antenna configurations and non-optimized CPICH power.

The simulated annealing algorithm used in the first part of the study and the Lagrangian heuristic for optimizing non-uniform pilot power are designed to handle large-scale planning scenarios. The average computing time of the two algorithms does not exceed fifteen and five minutes, respectively. The simulated annealing algorithm parameters have been defined as follows, $\delta_0 = 0.05$, $p_0 = 0.5$, $\delta_N = 0.01$, $p_N = 0.25$, $N = 2500$, $N_{bad} = 300$, and $maxTrials = 5$. Figure 5.3 demonstrates generated solutions, the algorithm moves, and the best found solutions in every fifth iteration when uniform CPICH power, antenna mechanical and electrical tilts, and antenna azimuth are optimized. The Lagrangian heuristic presented in Section 3.6 has been run with $N_1 = 500$, $N_2 = 50$, and $gap = 0.01$.

We make several observations from the results of uniform CPICH power in Table 5.2. Adjusting azimuth only gives a 34.5 % power saving in comparison to the reference value. Although this is a quite impressive amount, downtilting, in particular electrical downtilting, has an even larger effect. Note that downtilting leads to not only more significant power reduction, but also significantly smaller numbers of highly loaded and overloaded cells, as well as considerably lower DL load. When all three parameters are subject to optimization, the power reduction is 64.5 %, the number of highly loaded cells goes down from 24 to 3, and among them only one is overloaded. In average, DL load is improved by a factor of three. So far, we have not discussed results of cell overlap, which deserves a couple of remarks. First, optimizing all three configuration parameters reduces areas covered by four or more cells by 46.8 % — a significant improvement over the reference scenario. Second, the reduction is moderate for an overlap threshold of two cells, indicating that the negative impact on potential soft handover is small.

Let us examine solutions for non-uniform CPICH power. For the reference antenna configuration, optimization reduces the power of common channels by 53 % compared to the optimal uniform CPICH power solution. On the other hand, optimizing non-uniform power for the solution obtained from optimizing antenna azimuth and mechanical and electrical tilt gives a power reduction of 13 % only. Furthermore, the improvement in the other performance metrics is small when antenna configuration has been optimized for uniform power. Thus, the main benefit of adopting non-uniform power is less power consumption. In the last row of the table, power consumption of common channels is reduced to less than one third of that in the reference scenario. Another observation, which is probably more interesting than power reduction, is that combining electrical tilting and non-uniform CPICH power performs closely to the best results in the table. Thus most of the potential performance gain can be achieved at very lower cost!

Figures 5.2(a)-5.2(d) visualize coverage and load statistics of some network configurations. In the figures, we use the level of darkness to represent the number of overlapping cells. (In the darkest areas of Figure 5.2(a), ten cells overlap.) White areas are covered by one CPICH signal. The figures display also RBS locations and antenna azimuth. Each antenna is represented by a line. The line width reflects cell load and the length shows the antenna tilt. The longest lines represent antennas with zero tilt. Moreover, lines in red show overloaded cells, i.e., cells where some users are denied service. We observe that adopting non-uniform CPICH power, by itself, does reduce much overlap and cell load, in addition to power saving. The effects on reducing excessive cell overlap and improving cell load are clearly more dramatic if we optimize antenna configurations (see Figure 5.2(b)) or combine the two options (see Figure 5.2(d)). In comparison to Figure 5.2(a), antenna tilt and/or azimuth are changed in many cells in Figures 5.2(b) and 5.2(d), suggesting that a manual approach of tuning antenna configuration can hardly achieve the same performance as automated optimization.

5.6 Conclusions

Approaching service and performance goals in a UMTS network necessitates optimization of its design and configuration. In this chapter, we have addressed optimization of three RBS configuration parameters in UMTS networks: CPICH power, antenna tilt, and antenna azimuth. How these parameters are configured heavily influences service coverage, power consumption, and cell load. For this reason, the parameters are frequently used in tuning network performance. Two cornerstones of the presented approach are a system model that captures the interplay between the configuration parameters, and an algorithm that can deal with the complexity of the optimization problem.

Our case study demonstrates the benefits of the approach. Optimized CPICH power and antenna configuration offer significant performance gain. In comparison to the reference scenario, automated optimization reduces the power consumption of common channels and average cell load by approximately 70 %, and decreases the number of overloaded cells from 17 to only one. Furthermore, a significant reduction in CPICH power is a strong indication of reduced interference in the network.

Most of the performance gain can be conveniently implemented in the network through electrical tilting and adjusting CPICH power. Moreover, the presented algorithm is computationally efficient. It can be used as a tool to evaluate many potential solutions of network design (RBS location, sectorization, and antenna height) in an early stage of a network planning process.

There are a number of possible extensions of the work presented in this chapter. In particular, uplink coverage, which is an important performance factor, is not addressed in the presented work. Another interesting extension of the presented algorithmic framework is adopting it to HSDPA/HSUPA network planning.

Bibliography

[1] 3GPP TR 23.101. General UMTS architecture, v4.0.0. `http://www.3gpp.org`, Apr. 2001.

[2] 3GPP TR 25.401. UTRAN overall description, v4.6.0. `http://www.3gpp.org`, March 2003.

[3] 3GPP TS 25.101. User Equipment (UE) radio transmission and reception (FDD), v4.13.0. `http://www.3gpp.org`, Dec. 2006.

[4] 3GPP TS 25.104. Base Station (BS) Radio Transmission and Receiption (FDD), v4.9.0. `http://www.3gpp.org`, March 2007.

[5] 3GPP TS 25.133. Requirements for support of radio resourse management (FDD), v4.17.0. `http://www.3gpp.org`, March 2006.

[6] 3GPP TS 25.211. Physical channels and mapping of transport channels onto physical channels (FDD), v4.6.0. `http://www.3gpp.org`, Sep. 2002.

[7] 3GPP TS 25.304. User Equipment (UE) procedures in idle mode and procedures for cell reselection in connected mode, v4.8.0. `http://www.3gpp.org`, March 2004.

[8] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, New Jersey, 1993.

[9] E. Amaldi, P. Belotti, A. Capone, and F. Malucelli. Optimizing base station location and configuration in UMTS networks. *Annals of Operations Research*, 146(1):135–151, Sep. 2006.

[10] W. T. Araujo Lopes, G. Glionna, and M. S. de Alencar. Generation of 3D radiation patterns: A geometrical approach. In *Proc. of the 55th IEEE Vehicular Technology Conference (VTC2002-Spring)*, pages 741–744, May 2002.

[11] J. Borkowski, J. Niemelä, and J. Lempiäinen. Applicability of repeaters for hotspots in UMTS. White paper, Tampere University of Technology, Apr. 2005.

[12] P. Chitrapu, editor. *Wideband TDD : WCDMA for the Unpaired Spectrum*. Wiley, May 2004.

[13] A. Eisenblätter, A. Fügenschuh, E. R. Fledderus, H.-F. Geerdes, B. Heideck, D. Junglas, T. Koch, T. Kürner, and A. Martin. Mathematical methods for automatic optimisation of UMTS radio networks. Project report D4.3, IST-2000-28088 MOMENTUM, Sep. 2003.

[14] A. Eisenblätter, A. Fügenschuh, H.-F. Geerdes, D. Junglas, T. Koch, and A. Martin. Optimization methods for UMTS radio network planning. In *Proc. of the Intl. Conference on Operations Research*, pages 31–38, Sep. 2003.

[15] A. Eisenblätter, H.-F. Geerdes, T. Koch, A. Martin, and R. Wessäly. UMTS radio network evaluation and optimization beyond snapshots. ZIB-Report 04–15, Zuse Institute Berlin, Oct. 2004.

[16] A. Eisenblätter, H.-F. Geerdes, and N. Rochau. Analytical approximate load control in WCDMA radio networks. In *Proc. of the 62nd IEEE Vehicular Technology Conference (VTC2005-Fall)*, pages 1534–1538, Sep. 2005.

[17] A. Eisenblätter, T. Koch, A. Martin, T. Achterberg, A. Fügenschuh, A. Koster, O. Wegel, and R. Wessäly. Modelling feasible network configurations for UMTS. In G. Anandalingam and S. Raghavan, editors, *Telecommunications Network Design and Management*, Operations Research/Computer Science Interfaces, pages 1–24. Kluwer Academic Publishers, 2002.

[18] R. Esmailzadeh and M. Nakagawa. *TDD-CDMA for Wireless Communications*. Artech House Publishers, 2002.

[19] I. Forkel, M. Schinnenburg, and B. Wouters. Performance evaluation of soft handover in a realistic UMTS network. In *Proc. of the 57th IEEE Vehicular Technology Conference (VTC2003-Spring)*, pages 1979–1983, May 2003.

[20] M. Garcia-Lozano, S. Ruiz, and J. Olmos. CPICH power optimisation by means of simulated annealing in an UTRA-FDD environment. *Electronics Letters*, 23(39):2244–2247, Nov. 2003.

[21] M. Garcia-Lozano, S. Ruiz, and J. J. Olmos. UMTS optimum cell load balancing for inhomogeneous traffic patterns. In *Proc. of the 60th IEEE Vehicular Technology Conference (VTC2004-Fall)*, pages 909–913, Sep. 2004.

[22] A. Gerdenitsch, S. Jakl, Y. Y. Chong, and M. Toeltsch. A rule–based algorithm for common pilot channel and antenna tilt optimization in UMTS FDD networks. *ETRI Journal*, 26(5):437–442, Oct. 2004.

[23] F. Gil, A. R. Claro, J. M. Ferreira, C. Pardelinha, and L.M. Correia. A 3D interpolation method for base-station-antenna radiation patterns. *IEEE Antennas and Propagation Magazine*, 43(2):132–137, Apr. 2001.

[24] Global mobile Suppliers Association (GSA). `http://www.gsacom.com/`. Retrived in July 2007.

[25] A. Höglund. *Advanced mobile network monitoring and automated optimization methods*. PhD thesis, Helsinki University of Technology, March 2006.

[26] A. Höglund and K. Valkealahti. Automated optimization of key WCDMA parameters. *Wireless Communications and Mobile Computing*, 5(3):257–271, May 2005.

[27] H. Holma and A. Toskala, editors. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley & Sons, third edition, 2004.

[28] H. Holma and A. Toskala, editors. *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*. Wiley, July 2006.

[29] ILOG, Inc. Ilog Cplex *10.0 User's Manual*, Jan. 2006.

[30] T. Isotalo, J. Niemelä, J. Borkowski, and J. Lempiäinen. Impact of pilot pollution on SHO performance. In *Proc. of the 8th IEEE International Symposium on Wireless Personal Multimedia Communications (WPMC'05)*, Sep. 2005.

[31] IST-2000-28088 MOMENTUM. `http://momentum.zib.de`, 2003 (updated in 2005).

[32] ITU. IMT-2000 Radio Interface Specifications Approved in ITU Meeting in Helsinki. ITU Press Release, Nov. 1999.

[33] ITU (International Telecommuniction Union). `http://www.itu.int`.

[34] S. B. Jamaa, H. Dubreil, Z. Altman, and A. Ortega. Quality indicator matrices and their contribution to WCDMA network design. *IEEE Transactions on Vehicular Technology*, 54(3):1114–1121, May 2005.

[35] KATHREIN-Werke KG. `http://www.kathrein.de/`.

[36] D. Kim, Y. Chang, and J. W. Lee. Pilot power control and service coverage support in CDMA mobile systems. In *Proc. of the 49th IEEE Vehicular Technology Conference (VTC1999-Spring)*, July 1999.

[37] T. Kürner. Propagation models for macro–cells. Digital mobile radio towards future generation systems (COST 231 Final Report), COST Telecom Secretariat, CEC, Brussels/Belgium, Brussels, Belgium, 1999.

[38] J. Laiho, A. Wacker, and T. Novosad, editors. *Radio Network Planning and Optimisation for UMTS*. John Wiley & Sons, 2002.

[39] J. Lempiäinen and M. Manninen, editors. *UMTS Radio Network Planning, Optimization and QoS Management for Practical Engineering Tasks*. Springer, 2004.

[40] R. T. Love, K. A. Beshir, D. Schaeffer, and R. S. Nikides. A pilot optimization technique for CDMA cellular systems. In *Proc. of the 50th IEEE Vehicular Technology Conference (VTC1999 - Fall)*, volume 4, pages 2238–2242, Sep. 1999.

[41] R. Mathar and T. Niessen. Optimum positioning of base stations for cellular radio networks. *Wireless Networks*, 6(6):421–428, 2000.

[42] F. Mikas and P. Pechac. The 3D approximation of antenna radiation patterns. In *Proc. of the 12th Intl. Conference on Antennas and Propagation (ICAP 2003)*, volume 2, pages 751–754, March 2003.

[43] M. J. Nawrocki, M. Dohler, and A. H. Aghvami. On cost function analysis for antenna optimization in UMTS networks. In *Proc. of the 16th IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2005)*, pages 2327–2331, Sep. 2005.

[44] M. J. Nawrocki, M. Dohler, and A. H. Aghvami, editors. *Understanding UMTS Radio Network Modelling, Planning and Automated Optimisation: Theory and Practice*. John Wiley & Sons, 2006.

[45] J. Niemelä, T. Isotalo, and J. Lempiäinen. Optimum antenna downtilt angles for macrocellular WCDMA network. *EURASIP Journal on Wireless Communications and Networking*, 5(5):816–827, Oct. 2005.

[46] J. Niemelä and J. Lempiäinen. Mitigation of pilot pollution through base station antenna configuration in WCDMA. In *Proc. of the 60th IEEE Vehicular Technology Conference (VTC2004-Fall)*, pages 4270–4274, Sep. 2004.

[47] N. Papaoulakis, F. Casadevall, F. Adelantado, and E. Gkroutsiotis. Practical radio resource management techniques for UMTS. In *Proc. of Mobile Venue 2004*, 2004.

[48] J.-M. Picard, Z. Altman, S. Ben Jamaa, M. Demars, H. Dubreil, B. Fourestié, and A. Ortega. Automatic cell planning strategies for UMTS networks. *Intl. Journal of Mobile Network Design and Innovation*, 1(1):8–17, 2005.

[49] I. Siomina, P. Värbrand, and D. Yuan. An effective optimization algorithm for configuring radio base station antennas in UMTS networks. In *Proc. of the 64th IEEE Vehicular Technology Conference (VTC2006-Fall)*, pages 1–5, Sep. 2006.

[50] I. Siomina and D. Yuan. Optimization of pilot power for load balancing in WCDMA networks. In *Proc. of the 47th annual IEEE Global Telecommunications Conference (GLOBECOM 2004)*, volume 6, pages 3872–3876, Nov. 2004.

[51] I. Siomina and D. Yuan. Soft handover overhead control in pilot power management in WCDMA networks. In *Proc. of the 61st IEEE Vehicular Technology Conference (VTC2005-Spring)*, pages 1875–1879, May 2005.

[52] K. Tutschku. Demand-based radio network planning of cellular mobile communication systems. In *Proc. of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '98)*, pages 1054–1061, Apr. 1998.

[53] K. Valkealahti, A. Höglund, J. Pakkinen, and A. Flanagan. WCDMA common pilot power control with cost function minimization. In *Proc. of the 56th IEEE Vehicular Technology Conference (VTC2002-Fall)*, pages 2244–2247, 2002.

[54] K. Valkealahti, A. Höglund, J. Pakkinen, and A. Hämäläinen. WCDMA common pilot power control for load and coverage balancing. In *Proc. of the 13th IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002)*, pages 1412–1416, Sep. 2002.

[55] T. G. Vasiliadis, A. G. Dimitriou, and G. D. Sergiadis. A novel technique for the approximation of 3-D antenna radiation patterns. *IEEE Transactions on Antennas and Propagation*, 53(7):2212–2219, July 2005.

[56] L. A. Wolsey. *Integer Programming*. Wiley, 1998.

[57] J. Yang and J. Lin. Optimization of power management in a CDMA radio network. In *Proc. of the 52nd IEEE Vehicular Technology Conference (VTC2000-Fall)*, pages 2642–2647, Sep. 2000.

[58] S. Ying, F. Gunnarsson, and K. Hiltunen. CPICH power settings in irregular WCDMA macro cellular networks. In *Proc. of the 14th IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2003)*, pages 1176–1180, Sep. 2003.

[59] R. Zdunek, M. J. Nawrocki, M. Dohler, and A. H. Aghvami. Application of linear solvers to UMTS network optimization without and with smart antennas. In *Proc. of the 16th IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communiactions (PIMRC 2005)*, pages 2322–2326, Sep. 2005.

[60] H. Zhu, T. Buot, R. Nagaike, and S. Harmen. Load balancing in WCDMA systems by adjusting pilot power. In *Proc. of the 5th Intl. Symposium on Wireless Personal Multimedia Communications*, pages 936–940, Sep. 2002.

Part II

Coverage Planning and Radio Resource Optimization for Wireless LANs

Chapter 6

Introduction to Wireless LANs

Wireless Local Area Network (WLAN) is currently among the most important technologies for wireless broadband access. The flexibility offered by WLANs has been a major factor in their widespread deployment and popularity. Among the advantages brought by this technology are its maturity, low cost, and the ease of deployment of WLANs. The overall performance of a specific WLAN installation is largely determined by the network layout and its configuration. Among the necessary conditions for designing an efficient WLAN are therefore careful coverage planning and optimizing such network design parameters as access point (AP) locations, channel assignment, and AP transmit power allocation. These network planning and optimization tasks are in focus in the current part of the thesis. First, however, we provide a reader with some technical details of the WLAN technology and discuss the performance issues that need to be considered when planning such a network. This is the goal of the introductory chapter. In the next two chapters we present in detail our optimization approaches to WLAN planning addressing a number of performance aspects. In particular, in Chapter 7 we study the problem of optimizing AP locations and channel assignment taking into account coverage overlap of APs. In Chapter 8, our goal is to jointly optimize AP transmit power and channel assignment while minimizing potential contention among user terminals.

6.1 Technical Background

Wireless Local Area Networks (WLANs), nowadays often associated with Wi-Fi[1], are based on the IEEE 802.11 standard family[2] and have achieved tremendous popularity in recent years. The core of the IEEE 802.11 family is the standard known as "802.11 legacy" [1]. The standard specifies two raw data rates of 1 and 2 Mbps to be transmitted via infrared signals or by either Frequency-Hopping Spread Spectrum (FHSS) or Direct-Sequence Spread Spectrum (DSSS) in the Industrial Scientific Medical (ISM) frequency band at 2.4 GHz. For the medium access, the original standard defines two methods, the Distributed Coordination Function (DCF), a distributed random access scheme based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol, and Point Coordination Function (PCF), a collision-free centralized protocol. (The latter has been implemented only in a very few hardware devices, and it is not part of the Wi-Fi Alliance's inter-operability standard.)

[1] Wi-Fi, or Wireless Fidelity, is a brand originally licensed by the Wi-Fi Alliance, formerly Wireless Ethernet Compatibility Alliance (WECA), to describe the underlying technology of WLANs based on the IEEE 802.11 specifications. The charter for this non-profit organization was to perform testing, certify inter-operability of IEEE 802.11-based networking products, and promote the technology.

[2] IEEE 802.11 standard family is used to denote a set of WLAN standards developed by working group 11 of the IEEE (Institute of Electrical and Electronics Engineers) Standards Committee responsible for local area networks (LANs) and metropolitan area networks (MANs). The IEEE 802.11 standards are thus a part of a larger group of standards known as IEEE 802.

The first widely accepted wireless networking standard was IEEE 802.11b [4] followed later by IEEE 802.11a [3] and IEEE 802.11g [6]. In January 2004, IEEE announced that it had formed a new 802.11 Task Group (TGn) to develop a new amendment to the 802.11 standard for WLANs. The real data throughput is estimated to reach a theoretical 540 Mbps (which may require an even higher raw data rate at the physical layer), and should be up to 50 times faster than 802.11b, and up to 10 times faster than 802.11a or 802.11g. 802.11n builds upon previous 802.11 standards by adding MIMO (multiple-input multiple-output) that uses multiple transmitter and receiver antennas to allow for increased data throughput through spatial multiplexing and increased range by exploiting the spatial diversity. Since the standard is unfinished, it will not be further discussed here. The further discussions on the WLAN technology in this part of the thesis are applicable to any of the IEEE 802.11a/b/g standards unless the other is specified.

All IEEE 802.11 standards use the CSMA/CA medium access protocol, although differ by frequency spectrum, modulation schemes, and data rates. The IEEE 802.11a standard operates in the 5 GHz band with the total bandwidth of 20 MHz, whilst IEEE 802.11b and IEEE 802.11g operate in the 2.40 GHz band and divide the spectrum into 14 channels whose center frequencies are 5 MHz apart.

IEEE 802.11a uses a 52-subcarrier Orthogonal Frequency-Division Multiplexing (OFDM), and has a maximum raw data rate of 54 Mbps which, with an adaptive rate selection technique[3], depending on channel conditions, can be reduced to 48, 36, 24, 18, 12, 9, and then 6 Mbps. The supported modulation schemes are BPSK, QPSK, 16QAM, and 64QAM. IEEE 802.11b has a maximum raw data rate of 11 Mbps which can be scaled back to 5.5, 2, and then 1 Mbps. As its modulation technique, the standard uses the Complementary Code Keying (CCK), which is a slight variation on CDMA, extending the DSSS technology. IEEE 802.11g uses OFDM for the data rates supported by IEEE 802.11a (with the maximum data rate of 54 Mbps), but it can revert to IEEE 802.11b data rates with the corresponding modulation schemes if the signal quality becomes worse. Thus, although 802.11g operates in the same frequency band as 802.11b, it can achieve higher data rates because of its similarities to 802.11a and at the same time is backward compatible with IEEE 802.11b.

Being assigned a less "crowded" frequency band, IEEE 802.11a has the advantage of having more sub-carriers and experiencing less interference from, for example, microwave ovens, cordless telephones, Bluetooth, and other devices that operate in the frequency band assigned to IEEE 802.11b/g. On the other hand, due to higher frequencies, the signal attenuation is higher for IEEE 802.11a which makes the coverage range of the latter comparable, for example, to IEEE 802.11b (if the transmit power and other conditions are the same).

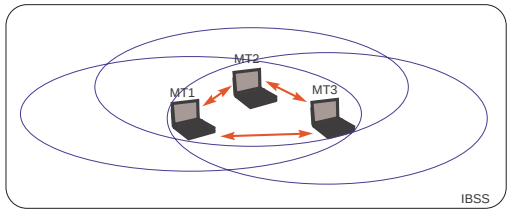
6.2 IEEE 802.11 WLAN Architecture

The IEEE 802.11 architecture consists of several components that interact to provide a WLAN that supports station mobility transparently to upper layers. This is a requirement in the IEEE 802 standard family [1] that the logical link control, the upper sub-layer of the OSI data link layer, must be the same for various physical media (e.g., Ethernet, token ring, and WLAN), and WLAN MAC must thus appear to the upper layers of the network as a "standard" 802 LAN.

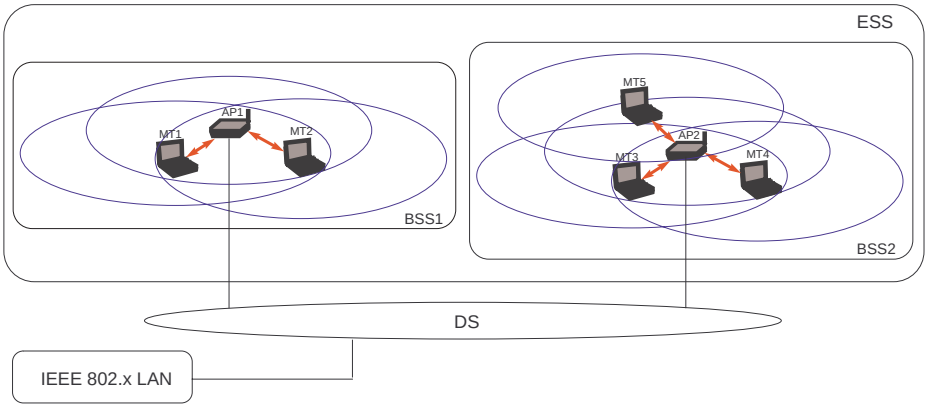
The WLAN architecture components are station, basic service set, access point, and distribution system. A *station* (STA) is a basic element of an IEEE 802.11 WLAN. It can be any device containing an IEEE 802.11 conformant media access control (MAC) and physical layer interface to the wireless medium.

A *basic service set* (BSS) is a basic building block of an IEEE 802.11 WLAN. A STA within a BSS can communicate with other members of the BSS, either directly or through some other STA.

[3]The link rate adaption algorithm has been left open in the standard.



(a) Ad hoc mode



(b) Infrastructure mode

Figure 6.1: IEEE 802.11 architectures.

A BSS within which STAs are able to communicate directly is called *independent basic service set* (IBSS). An IBSS can be viewed as a self-contained network. This is the most basic type of IEEE 802.11 WLAN. Because this network type is often formed without pre-planning, this mode of operation is often referred to as an *ad hoc network*.

In *infrastructure mode*, STAs in the same BSS do not communicate directly but through a special STA called *access point* (AP). We use the term "*terminal*" to refer to STAs without the AP functionality. In this mode, several BSSs may exist independently or may form a component of an extended form of network. In the second case, the architectural component used to interconnect BSSs is the *distribution system* (DS). In such an architecture, the closest communication point of any terminal is always an AP to which the terminal is associated, regardless of the destination address, and APs are connected to the DS. (Note that an IBSS has no access to the DS.) The DS enables mobility support and seamless integration of multiple BSSs transparently to upper layers. BSSs connected via the DS form the second type of IEEE 802.11 WLANs referred to as the *extended service set* (ESS) network. Stations within an ESS may communicate and move from one BSS to another, but to higher layers, the ESS appears the same as an IBSS network.

Figures 6.1(a) and 6.1(b) demonstrate an ad hoc network and a WLAN operating in the infrastructure mode, respectively. In Figure 6.1(a), three mobile terminals (MTs) act as STAs and form an IBSS. Since they belong to the same BSS and are within the communication range of each other (the communication range of each STA is marked with a solid line), any MT can directly communicate with the two others. In Figure 6.1(b), BSS1 and BSS2 are connected to the DS by AP1 and AP2, respectively. Moreover, through the DS, any MT from either of the two BSSs can communicate with devices that are a part of other IEEE 802 LAN (denoted in Figure 6.1(b) by IEEE 802.x LAN). Further in this part of the thesis only

WLANs operating in the infrastructure mode are discussed.

The recent trend both in industry and research community is a centralized WLAN architecture which is not a part of the IEEE 802.11 standard. In this architecture some functionalities (e.g., RRM, mobility management, QoS policies, security policies) are moved from APs into special devices that control a group of APs (see Section 6.5).

6.3 Media Access Control in IEEE 802.11 WLANs

For WLANs operating in the infrastructure mode, the IEEE 802.11 protocol defines two media access mechanisms implemented as the Distributed Coordination Function (DCF) and the Point Coordination Function (PCF). By the DCF, STAs negotiate medium access among themselves. An assignment regime (polling) is assured by an AP when the PCF function is activated. DCF is mandatory in IEEE 802.11 devices, whilst PCF is optional. Both mechanisms address two network performance issues, namely collisions and contention. *Collision* is a situation when one transmitter is overlaying another transmitter's signal if both devises attempt to transmit on the same channel at the same time. *Contention* is a situation when stations contend to access the shared medium. The PCF is a centralized mechanism which is designed to resolve both issues but suffers from bad scalability. The DCF is a distributed mechanism with is a trade-off between scalability and resolving the collision and contention issues.

The PCF is a polling-based protocol intended for sending time-sensitive information. With PCF, a point coordinator at the AP controls which terminals can transmit during a given period of time by polling them one at a time. This centralized MAC protocol can therefore only be used in infrastructure WLAN configurations. There is neither collision nor contention in a WLAN with a single AP using PCF. However, in a larger network, collisions can be completely avoided only when a centralized mechanism controlling all APs is present. This operation becomes very complicated and results in inefficient use of the allocated frequency spectrum when the number of APs increases. Since PCF has not been implemented on a large scale, we focus on DCF.

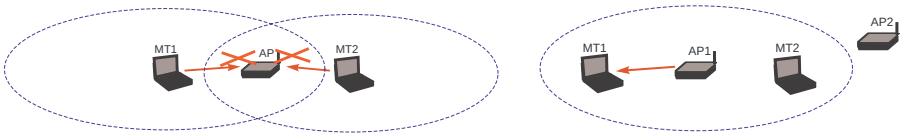
The DCF uses a random access scheme where each station has the right to initiate its transmission. This makes the scheme applicable not only in the infrastructure WLAN configurations but also in distributed and self-organized WLANs which has made DCF very popular. The DCF media access mechanism is based on the CSMA/CA protocol which extends the CSMA protocol.

CSMA is a contention-based protocol[4] making certain that all stations first *sense* the medium, i.e., try to detect the presence of an encoded signal from another station, before transmitting. This operation is known as *physical carrier sensing* and is performed at the physical layer. If a carrier is sensed busy, the station waits for the transmission in progress to finish before initiating its own transmission. By the IEEE 802.11 standard [2], physical carrier sensing is provided by the Clear Channel Assessment (CCA) function of the physical layer. Depending on the implementation, the CCA indicates a channel as busy when

- any received energy is above the energy detection threshold[5] (CCA Mode 1),
- a valid DSSS signal (either above or below the energy detection threshold) is detected (CCA Mode 2),
- a DSSS signal with energy above the energy detection threshold is detected, i.e., a combination of the two previous conditions occurs (CCA Mode 3).

[4] A protocol which is able to deal with contention.

[5] The IEEE 802.11 standard [2] requires the energy detection threshold to be at most $-80\,\text{dBm}$ for transmit power greater than $100\,\text{mW}$, at most $-76\,\text{dBm}$ for transmit power in the range of $(50, 100]\,\text{mW}$, and at most $-70\,\text{dBm}$ for transmit power below $50\,\text{mW}$.



(a) Hidden terminal problem (b) Exposed terminal problem

Figure 6.2: Hidden and exposed terminal problems.

Typically, the default energy detection threshold is much higher than the signal level at which a transmissions from a single station can be detected. Moreover, the threshold is usually adjustable only in higher-priced equipment. Assume that the interference comes either from own network or some other external IEEE 802.11b/g network(s). Then, the only situation when the CCA function in Mode 1 will report the medium to be busy and the CCA Mode 2 will not is when two co-channel transmissions collide and therefore cannot be recognized as valid DSSS signals. Such a situation does not happen very often which means that the second mode is more restrictive than the first mode, especially when the STA density is low. When the interference level is high, the two modes may become equally important. In the presence of strong non-DSSS signals in some parts of the network, the CCA in Mode 1 may report the medium to be busy more often than when Mode 2 is used. Therefore, the third option (CCA Mode 3) was introduced.

Since physical carrier sensing is used as the only means to avoid collisions in CSMA, the protocol suffers from collision and blocking problems. The two most common situations known as a hidden terminal problem and an exposed terminal problem are described below.

Hidden terminal problem. Terminal MT1 senses the carrier idle and starts transmission, and terminal MT2 does the same since they are outside the carrier sense (CS) range of each other. The result is a collision at the AP receiver (see Figure 6.2(a) where the terminal CS range is denoted by a dotted line).

Exposed terminal problem. A STA (terminal MT2) wants to initiate a transmission to another station (AP2) but is exposed to the ongoing transmission from the third station (AP1). The described situation is depicted in Figure 6.2(b).

The CSMA/CA protocol addresses the aforementioned problems arising with CSMA by reducing the collision and STA blocking probability through the use of an exponential back-off scheme, Inter-Frame Space (IFS), and virtual carrier sensing (in addition to the physical carrier sensing). Virtual carrier sensing is an optional mechanism which is not included in the CSMA/CA basic access scheme.

In CSMA/CA, the back-off scheme works as follows. At each packet transmission, if the channel is sensed idle, the STA transmits if the channel remains idle for a Distributed ISF (DISF). Otherwise, if the channel is sensed busy or gets busy during the DISF, the STA persists to sense the channel until it is sensed idle for a DISF and then generates a random back-off interval uniformly drawn from a range called *contention window*. The contention window size is doubled (within some valid range) after each unsuccessful transmission and reset after a successful transmission. The back-off counter is decremented as long as the channel is sensed idle, frozen when the channel is busy, and reactivated (without resetting) when the channel is again idle for more than a DISF. The transmission occurs when the back-off counter reaches zero. Observe that if the back-off counters of two STAs reach zero at the same time, a collision occurs. However, the probability of such a situation is significantly lower that with CSMA.

To prioritize acknowledgment packets, CSMA/CA uses a Short IFS (SIFS) which is delay of acknowledging a correct reception. Since the SIFS together with the propagation delay is

shorter than a DIFS, no other STA is able to detect the channel idle for a DIFS until the acknowledgment packet is sent. If the transmitting STA does not receive the acknowledgment packet within a specified time (SIFS + round trip time) or detects the transmission of a different packet on the channel, it starts the back-off counter according to the described back-off algorithm.

In addition to the basic access scheme which represents a two-way handshaking technique, DCF defines an optional four-way handshaking technique known as CSMA/CA with RTS/CTS. With this technique, prior the packet transmission a STA sends a special short Request-To-Send (RTS) frame. When the receiving STA detects an RTS frame, it responds, after a SIFS, with a Clear-To-Send (CTS) frame. The transmitting STA is allowed to transmit its packet only if the CTS frame is correctly received. Any other STA receiving the RTS frame, but not the CTS frame, is permitted to transmit (which solves the exposed terminal problem). Any STA, other than the intended destination, receiving the CTS frame should refrain from sending packets for a given time (which solves the hidden terminal problem). The amount of time the STA node should wait before trying to get access to the medium is included in both the RTS and the CTS frame. The RTS/CTS fame exchange is referred to as *virtual carrier sensing*. Observe that the exposed terminal problem and the hidden terminal problem are resolved only under the assumption that all stations have the same transmission range.

6.4 Performance Issues in IEEE 802.11 WLANs

Among the main advantages of WLANs that have made them so popular are their relatively low cost, the ease of deployment and network expansion, convenience, and mobility support. However, they have also some limitations. Some of them, e.g., short range, link instability, and low data rates, are mainly due to the underlying technology and therefore are very hard to tackle. The others (e.g., security issues, inefficient RRM, and low real throughput) are softer limitations that still have some room for improvement and are the tasks for system architecture designers, protocol designers, network planners, and networks administrators. Some of the important issues are discussed below. RRM strategies in IEEE 802.11 WLANs and related planning and optimization challenges are covered separately in Section 6.5.

Radio frequency interference. One of the major problems of IEEE 802.11 WLANs is that the network performance suffers a lot from interference that can be generated by STAs in own network, in some other IEEE 802.11 network operating in the same band, or by some other devices (e.g., microwave ovens).

- Because of the 802.11 medium access protocol, an interfering signal of sufficient amplitude and frequency can appear as a bogus 802.11 station transmitting a packet. This causes legitimate 802.11 stations to wait for indefinite periods of time until the interfering signal goes away.

- In another situation, when some device starts interfering while a legitimate 802.11 station is in the process of transmitting a packet, the destination will receive the packet with errors and will therefore not acknowledge the transmission which will cause the station to retransmit the packet adding overhead to the network. In some causes, 802.11 will attempt to continue operation in the presence of interference by automatically switching to a lower data rate, which slows the use of wireless applications by the transmitting user and also affects all other users in the same BSS (see also the performance anomaly effect discussed further in this section).

- Due to a small number of available non-overlapping channels, it is very likely that at least two neighboring APs operate on the same channel, especially when the network

spans over several floors. This may cause contention for the medium access between the APs and STAs in the corresponding BSSs. To resolve the contention issue, overlapping channels may be used. This, however, may increase interference and in some cases may have an even more harmful effect on the network performance than using a small number of channels (e.g., when the STA density is high [19]).

A consequence of increasing WLAN deployment, coupled with limited number of channels and unlicensed spectrum usage, is that the interference between transmissions is becoming a serious problem.

Unbalanced UL and DL contention. The DCF does not distinguish between APs and terminals which makes APs competing for the medium access on equal terms with terminals. This results in an unbalanced contention between UL and DL traffic and may result in a situation when contention caused by intensive UL traffic load starves the DL traffic of an AP. In other words, the AP becomes a network bottleneck.

Low net throughput. It has been proven analytically and demonstrated by simulations that the maximum throughput that a user can expect in a WLAN with DCF is significantly lower than the nominal bit rate due to the CSMA/CA protocol characteristics and the amount of overhead bits. For example, the author of [30] analytically showed that, even with no collisions, a single station sending long frames with a payload of 1500 bytes and MAC Protocol Data Unit (MPDU) of 1534 bytes will have a maximum useful throughput of 70 % (7.74 Mbps). Modeling and measurement results in [41] showed that net throughput values for TCP traffic in an IEEE 802.11b WLAN are of respectively 82 %, 76 %, 62 %, and 47 % at 1, 2, 5.5, and 11 Mbps. For an IEEE 802.11a network, the same authors found the net TCP throughput of 80 % at 6 Mbps and 55 % at 54 Mbps. Similar effects have been observed in Section 7.3.

Short-term unfairness. The goal of the CSMA/CA protocol is to provide fair equal-probability medium access in a long term for all stations. This, however, does not guarantee short-term fairness in the network [45]. The explanation is that a situation, when a small number of stations starve the others, has significant performance implications for applications and transport protocols, e.g., for low jitter in real-time audio and video applications.

Performance anomaly. When a station captures the channel for a long time because its bit rate is low, it penalizes other stations that use a higher bit rate [30]. The situation is a common case in WLANs in which a station far away from an AP is subject to significant signal fading and interference. To cope with the problem, the station changes its modulation type which results in a lower bit rate degrading the throughput for other stations. The problem is closely related to short-term unfairness.

Non-controllable MT transmit power. The MT's transmit power in IEEE 802.11 WLANs is not controlled by APs. Every MT equipment can choose any transmit power level from a predefined (discrete) set. However, being not restricted by the serving AP, the MT usually transmits at the highest possible power level unless the transmit power is manually set. This strategy has some reasons (e.g., getting higher data rates or having more potential servers), but this is a short-sighted view since this may have another effect — increased throughput for other users but no significant improvement for the current user. Furthermore, the maximum transmit power of APs can be significantly lower than that of MTs (compare, for example, [16] and [18] where the maximum transmit power level of an AP and an MT is 30 mW and 100 mW, respectively). This results not only in high interference generated by MTs and high contention in the network, but also makes RRM techniques based on AP transmit power

adjustment less efficient. Also, recall that the RTS/CTS mechanism does not completely resolve the hidden terminal problem and the exposed terminal problem (see Section 6.3).

A lot of research has been conducted to address the aforementioned issues. Most of the proposed remedies to the aforementioned issues require significant hardware and/or software upgrades at either the MT or the AP side or both. The examples are the network performance enhancements through the use of directional antennas with multiple receivers [68], optimizing the contention window size and the binary exponential back-off procedure (see, for example, [54]), more sophisticated rate adaption techniques (see for example, [48]), power control algorithms (in, for example, [33]), etc.. Efficient network planning allows to partially solve or address most of the mentioned performance issues and significantly improve the network performance at a relatively low cost.

6.5 Network Planning and RRM Challenges in IEEE 802.11 WLANs

As it follows from Section 6.4, the performance of IEEE 802.11 WLANs is far from being perfect nowadays and therefore, has a lot of space for improvement. In this context, developing efficient RRM schemes has been attracting attention of many researches (see Section 6.6) and standardization bodies (see Section 6.5.5). This trend is due to the growing popularity of WLANs and the interest in deploying them on a large scale, on one side, and due to the lack of efficient radio resource management mechanisms for IEEE 802.11 networks, on the other side. As another trend, a lot of research has been dedicated to developing efficient automatic planning tools for WLANs (see, for example, [39]). This is mainly because of relatively low prices for the equipment and the ease of WLAN deployment which makes it unnatural to invest a lot of efforts and money in the network planning process due to a high flexibility in changing the network topology, usually smaller network deployment and operation budgets (as compared, for example, to cellular networks), and very often non-commercial purpose of WLANs (corporate networks, for example).

The main goals of efficient radio resource planning for IEEE 802.11 networks are reducing contention and interference, providing at the same time good coverage and ensuring high throughput. With respect to these goals, the main focus of RRM is primarily on developing efficient power control, automatic channel assignment, and load sharing mechanisms between APs. This, however, is a challenging task due to the CSMA/CA mechanism and a very dynamic nature of radio propagation conditions in indoor environments. Also, the importance of finding optimal AP locations should not be underestimated when planning the WLAN topology.

6.5.1 Channel Assignment

The IEEE 802.11b and 802.11g standards divide the 2.4 GHz spectrum into 13 channels separated by 5 MHz. In addition to this, IEEE 802.11b specifies one more channel for Japan. Channel availability varies across geographical regions due to different spectrum regulations. For example, FCC (Federal Communications Commission) and IC (Industry Canada) restrict the spectrum usage in North America to only 11 channels. 13 channels are typically available in the ETSI (European Telecommunications Standards Institute) regulatory domain. In Spain and France, only two and four channels, respectively, were originally allowed by the regulatory bodies for the two standards [4, 6]. However, the available spectrum has been later extended to 13 channels in both countries. In Japan, 14 and 13 channels can be used for IEEE 802.11b and IEEE 802.11g, respectively.

In addition to the center frequencies, the standard specifies a power envelop by which the signal must drop by at least 30 dB below peak energy at ± 11 MHz and by at least 50 dB at ± 22 MHz from the center frequency. Channels at least 24 MHz apart are often

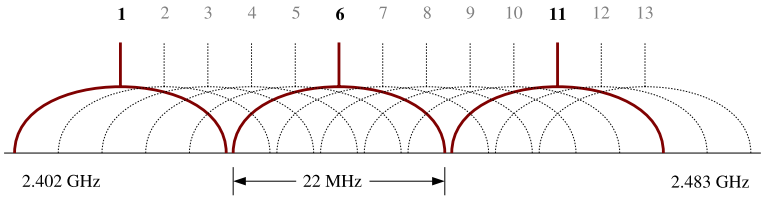


Figure 6.3: Three non-overlapping channels specified by the IEEE 802.11b/g standards.

considered to be *non-overlapping* (see Figure 6.3). This yields at most three non-overlapping channels. Channels with maximum spacing of ±25 MHz in the 2.4 GHz spectrum are defined as *adjacent channels* [4]. Often, these channels are also called *overlapping channels*. In practice, however, channels with an even larger spacing may still overlap, i.e., interfere with each other [9, 24]. For example, it has been demonstrated by measurements that a powerful transmitter operating on channel 1 can effectively interfere those operating on channel 6 or even channel 11.

When planning 802.11 WLANs, using non-overlapping channels (typically, channels 1, 6, and 11) is usually encouraged [19] since STAs transmitting on non-overlapping channels do not contend for the medium. Therefore, channel assignment for IEEE 802.11g Wireless LANs is usually understood as a frequency assignment problem with three available frequencies. However, this view is simplified for the following reasons. First, this approach does not take into account interference that still may be generated on channels commonly known as non-overlapping. Therefore, assigning channels 1 and 6 to two neighboring APs, for example, is considered as good as assigning channels 1 and 13. Second, three channels are insufficient for high-density WLANs [26]. Third, the total number of available channels varies by country. Furthermore, the set of available channels can be further reduced in some local environments. For example, the set of possible channels in some parts of the network can be negotiated with the administrator of a neighboring network. Sometimes, a network planner may also want to avoid using some channels at certain APs to avoid interference coming from neighboring WLANs or other sources (to detect which a careful RF site survey is necessary). It has been also shown by simulations that using partially overlapping channels can improve network performance [56], but if adjacent channel interference is carefully planned.

For the IEEE 802.11a standard, up to 12 non-overlapping channels are available which seems to simplify the channel assignment task for this type of WLANs due to missing adjacent channel interference. However, receivers of many cheaper WLAN adapters cannot cleanly filter out single channels. As a result, they experience interference from adjacent channels as well [67], although, clearly, adjacent channel interference is usually a smaller issue in IEEE 802.11a WLANs due to a larger number of available channels and better separation.

To the best of the author's knowledge, distributed dynamic channel assignment is not implemented in today's APs. Typically, a WLAN either uses a static channel assignment (the assignment is usually done manually) or the APs use simple heuristics such as searching for the least congested channel, e.g., the one with the least amount of traffic or with the least amount of interference. The procedure may be either performed on request or triggered automatically at a specified interval (for example, one hour). The main disadvantage of this approach is that the resulting network configuration can very often suffer from the hidden terminal problem, beside that the found channel assignment is not globally optimal even with respect to the considered metrics.

One of the main difficulties of distributed dynamic channel assignment is that the decision cannot be taken at each AP separately without a risk of network performance degradation. To overcome this, intensive information exchange is required between neighboring APs, which brings new, not yet fully solved issues like, for example, communication between APs,

synchronization, and communication overhead. To facilitate communication between APs, IEEE 802.11F proposes Inter Access Point Protocol, but it does not provide sufficient support for RRM-related communications. The IEEE 802.11k standard, which is not complete yet, aims to facilitate RRM in WLANs but it is unclear if distributed channel assignment algorithms can exploit this (see Section 6.5.5 for more details). Inter-AP communication via the AP radio interfaces will affect a lot user throughput. Moreover, data transmissions can be disturbed while APs are switching to a different channel. Thus, asynchronous channel switching among APs is inefficient from the network performance point of view. Because of the aforementioned issues, centralized approaches to RRM has gained attention from both academic and industrial researches (see Section 6.5.4).

6.5.2 Transmit Power Control

Transmit power control is a mechanism of automatic adjustment of transmit power level according to a current situation (e.g., interference). This is an important technique widely used in wireless networks to optimize capacity and reduce power consumption. In cellular networks, power control is a dynamic process that enables power adjustment with very high frequency. In WLANs, however, the concept of power control still remains an open issue due to a complex effect of adjustable power levels on overall network performance. (So far, transmit power control has been defined only for IEEE 802.11a networks [7] with the purpose to keep the interference below a given maximum.) Moreover, it is still unclear how dynamic power control in WLANs should be. Therefore, power control in WLANs is, as usually, understood as a very frequent power adjustment, although it still can be viewed as a periodic (with significantly longer time intervals) power adaption based on the collected statistics. The latter is particularly common for centralized architectures.

AP transmit power control in IEEE 802.11b/g networks could be utilized for controlling coverage, load sharing between BSSs, and reducing interference (see, for example, [15, 27, 33, 53]). There exist two views on power control in 802.11b/g WLANs [53]. First, the concept can be applied to transmit power in a usual sense, allowing a STA to spend no more than the amount of power needed to reach the AP they are associated to. This view, however, overlooks the hidden terminal problem which will arise due to different transmit ranges at neighboring STAs [33]. To address this issue, it has been suggested to also adjust the energy detection threshold [53]. Another issue with the transmit power control is that it involves a trade-off between interference and contention reduction, on one side, and the rate adaption strategy, on the other [11]. Other difficulties of implementing power control in WLANs in a way similar to cellular networks are large channel quality variation in indoor environments and the access scheme defined by the CSMA/CA protocol which will make the network performance suffer a lot from frequent transmissions of control information.

The second view is to adjust transmit power when transmitting special frames. Since the decision on which AP is to be selected is taken by MTs, the IEEE 802.11k standard will allow an AP to control the extent of its coverage area by selecting its transmit power of certain packets (beacon frames and probe response frames) that are used by MTs for deciding their association [53, 27]. In a situation when an MT is forced to choose another AP on the same channel, the effect is straightforward (provided that the MT does not change its transmit power). However, if with the same strategy, the MT will have to change the channel or/and increase its transmit power, the MT may degrade the service quality even in its old BSS (e.g., when it becomes a strong interferer to neighboring STAs and transmits in parallel) or unpredictably increase contention in the network.

6.5.3 AP Placement

Selecting AP locations is the first step in the WLAN deployment process and it has to be done such that the designed WLAN could provide complete coverage of the intended area.

WLAN topology design is usually based on signal strength measurements and consideration of radio propagation issues. This, however, can be a challenging task, especially when the WLAN is planned over multiple floors and therefore, an AP located on one floor of the building may provide signal coverage to adjacent floors of the same building or to another building as well [31, 32]. A straightforward strategy when selecting AP locations with the main focus on coverage planning is to space them as far apart as possible while still providing the necessary coverage. This simple strategy allows to minimize the equipment costs and to reduce potential coverage overlap of APs that later may be assigned the same channel. Careful RF site survey may also be very helpful in detecting possible sources of interference and avoiding placing APs in problematic areas.

6.5.4 Automated Radio Network Planning and RRM

Although some of the WLAN planning tasks require direct human involvement, e.g., RF site survey or defining candidate AP locations, most of the radio network planning and RRM routines could be effectively automated. This is especially desirable when designing and optimizing large networks, i.e., when manual network configuration becomes a tedious task.

Automated Planning for a Static Configuration in WLANs

Planning WLAN topology and initial network configuration relates to static planning and usually implies finding optimal AP locations and defining optimal channel assignment such that full coverage of the target area is achieved, and contention probability and interference are minimized. The coverage plan is typically designed assuming the maximum transmission range of APs. In some cases, however, AP transmit power can also be decided, e.g., to control the coverage area of some APs or to balance the load among APs when traffic distribution is known.

In practice, these decisions are usually taken one in each step, i.e., first, AP locations are decided, and then channels and (less often) AP transmit powers are defined [32]. The decisions are mostly based on network planners' experience. Instead, if the radio propagation environment can be properly modeled, optimization techniques can be utilized to find an optimal network plan with respect to a given objective and subject to the defined constraints. The optimal network plan will then give the locations where the APs are to be installed (as a subset of given candidate AP locations), channel setting, and AP transmit power levels (if this is a part of the network planning task). Such an approach would make the planning process more efficient for at least the following reasons. First, the process can be automated which will reduce the amount of manual work and will make it possible to experiment with different virtual network plans before deploying the real network. Second, simultaneous optimization of several decisions allows to avoid suboptimal solutions.

The advantage of the optimization approach is demonstrated in Chapters 7 and 8. In Chapter 7, several optimization models addressing the AP placement and channel assignment problems are proposed. The models can be effectively used for static WLAN planning. The optimized network plans are compared to those obtained by a traditional sequential planning approach. In Chapter 8, the proposed model also allows us to decide the AP transmit power levels.

Automated Radio Resource Control

Once the network is deployed, the initial network configuration, even being optimal with respect to the static planning goals, can be further adapted to dynamic radio propagation conditions, user distribution, and traffic load. Among the control parameters, there are channel setting, transmit power level, energy detection threshold, etc.. One might also think of a dynamic topology in which APs can be dynamically switched on when they are needed and switched off, otherwise.

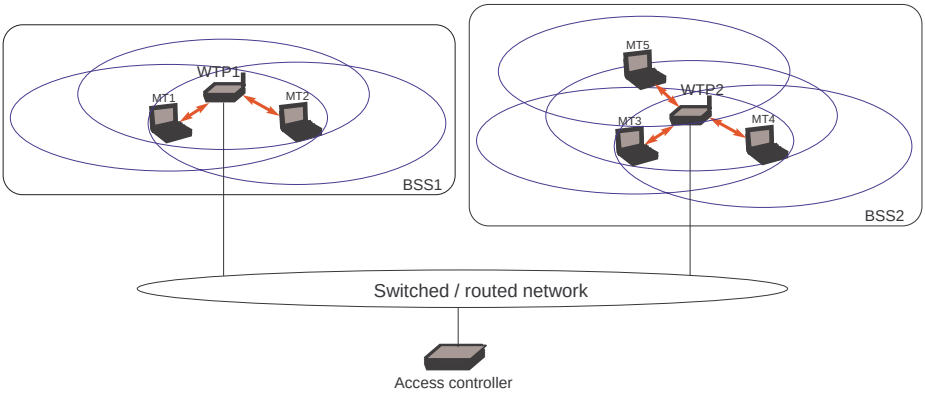


Figure 6.4: Centralized WLAN architecture.

The dynamic aspect of radio resource control is particularly important in WLANs due to the behavior of the CSMA/CA protocol and indoor environments (where WLANs are most often deployed) which cause very frequent variations of radio propagation characteristics and may tremendously affect connection speeds and error rates. This makes dynamic radio resource control in WLANs necessary. However, for the aforementioned reason and because of the issues discussed in Sections 6.5.1 and 6.5.2, the frequency of performing dynamic resource reallocation is decided by a trade-off between the control overhead, network stability, and configuration optimality.

Analyzing the aforementioned trade-off, several conclusions can be drawn. First, the RRM mechanisms in WLANs must be rather statistical than deterministic. Second, resource reallocation (e.g., channel assignment) cannot be performed very often to ensure network stability, which makes using off-line computations acceptable. Third, although distributed architectures provide more flexibility and are characterized by better scalability, distributed RRM involves a lot of control information overhead [25] which may significantly affect the final user throughput. Fourth, the effect of changing the channel assignment and transmit power is complex and is very difficult to predict in a distributed manner in large networks where APs are spread in a three-dimensional space [32, 53]. All these statements suggest that optimization techniques can be useful not only for static planning but also for dynamic RRM, especially in a centralized architecture which has recently become a popular trend in industry.

The advantage of a centralized architecture is that it can ensure a consistent RRM strategy in the controlled area and thus resolve such issues as conflicting RRM decisions (that are very likely to happen due to, for example, small number of available channels) taken by single APs, excessive inter-AP communications, unfairness among BSSs, etc.. For larger networks, a fully centralized architecture may be not very efficient, although it can be adapted and utilized on a cluster level in a cluster-like architecture. In most of the today's WLANs, however, the centralized approach will probably work very well. Therefore, even not being directly supported by the IEEE 802.11 standard, the centralized architecture approach has been widely adopted by industry (see, for example, [20]).

In the centralized architecture, most functionalities, including RRM, are moved from APs to another device (e.g., intelligent switch, WLAN controller, access controller, etc.). The APs are therefore referred to as Lightweight APs (LAPs). By one of the approaches called "split MAC" architecture, the LAPs are to be kept as light as possible such that they could act only as a radio interface between MTs and the new central device, and all the IEEE 802.11 MAC functions are to be moved to the new devices that communicate with LAPs via a special

protocol [20]. The maximum number of LAPs that can be supported by one control device varies in the range from 6 to 100, depending on the manufacture and the model.

The centralized WLAN architecture is described in detail in [66] where Wireless Termination Points (WTPs) act as LAPs and are controlled by one or more access controllers to enable network-wide monitoring, improve management scalability, and facilitate dynamic configurability. The WTPs are connected to access controllers via an interconnection medium which is a switched or routed network (see Figure 6.4). The idea of introducing special devices (intelligent switches) into the WLAN architecture was also discussed in [32] where the authors describe possible dynamic channel assignment, AP transmit power control, and load sharing strategies for the proposed architecture.

6.5.5 Standardization Work towards Efficient RRM

Although WLANs have been already quite some time on the market and have experienced great popularity, only simple RRM features have been so far provided in the IEEE 802.11 equipment. One of the reasons is that the major part of the RRM functionality in WLANs has been left so far open by the standard and thus decided by equipment manufactures. Because of the insufficient standardization support for RRM, it is even more important to know what is currently covered by the standards. The goal of this section is to present the finalized standards and to also sketch ongoing standardization efforts that aim at enabling efficient RRM in IEEE 802.11 WLANs.

Please note that although that the IEEE 802.11 standard family, as it is in present, does not directly support the centralized architecture that has recently become very popular (see Section 6.5.4), the new RRM-related standards are expected to work both in distributed and centralized architectures.

IEEE 802.11k (Radio Resource Management Enhancements) is a proposed standard for radio resource management that defines and exposes radio and network information, i.e., a set of measurements and reports (for example, roaming decisions, RF channel knowledge, hidden STAs, client statistics, transmit power control, etc.) to be exchanged between APs and MTs in order to facilitate radio resource management during operation and maintenance of mobile WLANs. 802.11k will allow MTs to discover the best available AP not only based on the received signal strength (as it is now) but also with respect to the AP capacity and the current load. The standard has not been finalized yet.

IEEE 802.11v (Wireless Network Management). A recently started work on the IEEE 802.11v standard aims at developing procedures enabling efficient load balancing solutions. The standard will allow management (monitoring, configuring, updating) of client devices through layer 2 while they are connected to IEEE 802.11. The standard will probably include cellular-like management paradigms and will compliment 802.11k that gathers information from the stations.

IEEE 802.11F (Inter-Access Point Protocol) is a recommendation that describes an optional extension to IEEE 802.11 that provides wireless communication between APs from different vendors [5]. IEEE 802.11F proposes Inter-Access Point Protocol (IAPP) intended to facilitate and speed up handover within an ESS but does not provide sufficient support for RRM-related communications.

IEEE 802.11h (Spectrum and Transmit Power Management Extensions in the 5 GHz band in Europe) is the standard for spectrum and transmit power management extensions that solve problems like interference with satellites and radar using the same 5 GHz frequency band [7]. Originally designed to address European regulations, it is now

applicable in many other countries. The standard provides Dynamic Frequency Selection (DFS) and Transmit Power Control (TPC) functionalities to the IEEE 802.11a MAC. DFS aims at avoiding co-channel operation with radar systems and ensuring uniform utilization of available channels. TPC ensures that the average power is less than the regulatory maximum in order to reduce interference with satellite services.

6.6 Related Work

Much research has been done for planning mobile telecommunication networks (see Chapter 1.1 for more details). One planning aspect in this context is selection of installation sites from a set of available candidate sites (see, for example, [29]). This can be combined with configuring of RBSs, e.g., choosing antenna types, orientation, and transmit power [34]. Often, the placement problem has multiple, competing objectives, such as maximizing coverage, maximizing capacity, and minimizing installation cost. The latter, however, is more important, for example, for UMTS networks, and it is less important for WLANs due to relatively low equipment costs. The planning aspects we will address are AP placement, channel assignment, and AP transmit power. In many research works, these planning targets are considered separately and sometimes viewed as three separate sequential phases of the network planning process (see, for example, [31]). A similar approach was applied by Wertz et al. [65]; the authors proposed several greedy strategies for first finding AP locations and subsequently assigning channels, and presented results for a realistic indoor planning scenario.

Optimizing AP locations alone was studied by Kamenetsky and Unbehaun [40] with the goals of maximizing the coverage area and the overall signal quality by using a convex combination of two objectives (the approach is also used in [60] and has been adopted in Section 7.2.3). One of the objectives aims at improving the average signal quality (minimizing the average path loss) over the entire service area, and the second objective minimizes the areas with poor signal quality. The authors compared different solutions obtained by several heuristics (pruning algorithm, neighborhood search, and simulated annealing) and studied the algorithm performance. The model did not consider any characteristic specific to WLANs, such as carrier sensing or contention, and therefore, can be applied for planning other radio networks as well.

Amaldi et al. [12] defined WLAN efficiency as the sum of probabilities of accessing the network over all users and showed that the problem of designing WLANs with maximum efficiency gives rise to extensions of the classical set covering problem in which the AP coverage overlap is minimized. The authors provided the exact formulation of the problem (a generalization of the 0-1 hyperbolic sum problem) which is $\mathcal{NP}$-hard. Assuming that in the optimal solution the number of covering APs in each test point is small and limited by a constant, the authors derived effective linearizations that allow to tackle large instances. The authors considered only a single-channel scenario, but claimed that the proposed model can be easily generalized to account for multiple channels. To efficiently solve the problem utilizing the obtained linearization results, the authors of [13] proposed heuristics combining greedy and local search phases and presented a numerical study for a real office environment.

Channel assignment (also known as frequency planning) has been extensively studied for other technologies [8], notably GSM. Many frequency assignment models known for cellular networks can be often viewed as graph coloring problems [23]. However, not all the models can be reused for WLANs. For example, one of the most common objectives in frequency assignment models for cellular networks is to minimize the frequency span, which does not make sense for WLANs where the whole frequency band is freely available. Another type of well-studied frequency assignment problems, feasible frequency assignment, aims at finding a feasible assignment satisfying certain constraints that usually ensure the minimum channel distance between neighboring cells. With a high STA density, ensuring these constraints will result in no feasible solution.

Among the well-known frequency assignment problems, the most straightforward and probably the easiest to adapt to WLAN planning seems to be the minimum interference frequency assignment problem that minimizes the total interference induced by channel overlap. In this model, adjacent channel interference can be modeled using weighting factors that depend on the channel distance. A simplified modeling approach here is based on the assumption that only non-overlapping channels can be used (three channels for 802.11g networks), i.e., the corresponding weighting factors are set to zero.

Mishra et al. [55] formulated a weighted variant of the graph coloring problem specifically designed to allow the use of overlapping channels and to take into account realistic channel interference. The interference on adjacent channel is modeled by weights defined experimentally. The authors showed that the problem is $\mathcal{NP}$-hard and proposed two distributed heuristic algorithms.

A channel assignment model that takes into account the 802.11 medium access aspects was proposed by Leung and Kim in [50]. The authors defined the effective channel utilization at an AP as the fraction of time during which the channel can be sensed busy or used for transmission by the AP, and formulated the problem that minimizes the effective channel utilization at the bottleneck AP. The authors proved that the problem is $\mathcal{NP}$-hard and presented a heuristic approach, which they applied to two scenarios with known optimal channel assignments and uniform fixed-power sectorized antennas.

Riihijarvi et al. [57] and Garicia Villegas [26] discussed implementation issues of dynamic channel assignment in real wireless networks. In [57], the authors demonstrated how graph coloring can be used as a theoretical basis for designing a protocol. In [26], an algorithm for distributed dynamic channel assignment maximizing the cell utilization was presented, and details on its implementation for a real network were discussed. Dynamic channel assignment was converted into a graph coloring problem also by Hills and Friday in [32]; the authors considered minimizing co-channel overlap and suggested to use inter-AP signal strengths as overlap measures.

AP placement and channel assignment are jointly treated in some works. However, channel assignment has in general only been addressed in the form of a feasibility problem (see, for example, [49, 58]). Rodrigues et al. [58] presented an integer programming model that improves coverage by maximizing the total signal strength over the area taking also into account the importance of each TP. By the model, the minimum channel distance must be maintained for each pair of interfering APs, which causes infeasibility if co-channel interference cannot be avoided. The authors used the propagation data obtained from real measurements and solved the problem by Ilog Cplex [63].

Lee et al. [49] presented an integer programming formulation that minimizes the maximum of channel load, or channel utilization, at APs, while satisfying the traffic demand. The authors aimed at finding the best AP locations with non-overlapping channels such that the available bandwidth is not exceeded and the full coverage of the area is maintained. The channel utilization is modeled as the portion of the bandwidth allocated to user traffic. The problem was solved by Ilog Cplex [63] for up to 40 candidate AP locations. The disadvantage of the model is that it does not provide any feasible solution if the interfering APs cannot be assigned non-overlapping channels. The authors also formulated a network reconfiguration problem that minimizes the traffic demand disruption due to reassignment of demand points and a new channel assignment such that the maximum channel utilization does not exceed some value. The traffic disruption is defined as the sum of added and subtracted traffic flows (with respect to the previous traffic amount).

I a few works AP placement and channel assignment are jointly optimized. For example, Ling et al. in [51] presented the patching algorithm that maximizes the total network throughput scaled with Jain's fairness index [36] defined with respect to individual user throughput measures. The individual user throughput is defined as the product of data rate, probability that the user holds the channel, and the portion of time used by the user for transmitting its

payload. The algorithm successively places a given number of APs, one by one, and in each step sequentially tests all possible channels at each of the selected APs in order to improve the objective. Only co-channel overlaps were considered by the authors.

The problem of modeling and minimization of contention under CSMA/CA-type protocols is addressed by Zdarsky et al. [67]. The authors presented a model that has a polynomial (non-linear) structure. Exploiting the binary nature of most variables, the model was then transformed to an equivalent linear model. The model provides decisions on MTs' association, transmit power levels of all STAs, and channel assignment. Only non-overlapping channels are assumed. The authors used a mixed integer programming solver (open-source software `lp_solve` [62]) to find optimal solutions for small networks (4 APs, 5 MTs, up to 4 channels), but had to resort to a custom genetic algorithm heuristic to deal with larger instances (200 APs, 400 MTs, 4 channels; 100 APs, 500 MTs, 4 channels). The authors derived also theoretical lower bounds on the minimum contention for low and high traffic scenarios, taking also into account contention induced by the RTS/CTS mechanism. The lower bounds are very tight for small test networks, but increase significantly with the problem size giving a gap of 50.3 % and 31.8 % for the first large test network for low and high traffic, respectively, and a gap of 119.9 % and 15.87 % for the second large test network for low and high traffic, respectively.

Jaffrès-Runser et al. [35] recently presented a multi-objective optimization algorithm that simultaneously maximizes coverage, minimizes co-channel interference, and maximizes a QoS metric, with AP locations and AP transmit power levels as decision variables. The interference metric at each TP reflects the signal strength coming from the strongest co-channel interferer. Provided that h is the number of non-overlapping channels, the strongest co-channel interferer for each TP is defined as the AP at position $h + 1$ in the sequence of APs sorted in descending order by their path gains. The CSMA/CA protocol functionality and rate adaption are captured in the QoS metric which takes into account effective user throughput [52]. The algorithm is based on Tabu search and exploits the Pareto optimality criterion. A neighboring solution is obtained by either moving one candidate AP to a new position, adding a new AP, or changing transmit power of an AP. In each iteration, the algorithm selects a set of non-dominated solutions from a neighborhood of the current search front. The set is used to update the optimal front and to derive a subset of solutions to be added to the current search front. Tabu list contains members of a search front whose neighbors were included in the current search front at some iteration. The channel assignment is proposed as the next planning step, i.e., solved separately from the AP placement and transmit power assignment problem. The authors claimed that the interference metric implicitly addresses channel assignment and simplifies a-posteriori channel assignment resolution.

Chapter 7

Optimization of AP Locations and Channel Assignment

Given a fixed amount of AP equipment and a set of available channels, we aim at finding an efficient design of a Wireless LAN with respect to a channel overlap measure and the expected net user throughput over the wireless service area. Although the presented approach and the system model are applicable to any type of IEEE 802.11 network operating in the infrastructure mode, in our experiments we focus on IEEE 802.11g networks since they are typically used in office environments.

7.1 System Model

Let $\mathcal{A}'$ be a set of candidate AP locations. The candidate locations are determined in advance with respect to different factors such as potential installation costs, accessibility, physical security, radio propagation aspects, health safety, psychological factor (e.g., not all people are willing to have an AP installed in their office). In problems where AP locations are a part of the decision to be taken, the maximum number of APs to be installed is restricted by M. The set of selected APs is denoted by $\mathcal{A}$ (this set is used as input when solving channel assignment problems).

If installed, AP a transmits at a power level from a predefined discrete set $\mathcal{L}$ and uses a channel from a set of available channels $\mathcal{C}$. We use $\mathcal{D}$ to denote the set of possible channel distances, i.e., $\mathcal{D} = \{|c_1 - c_2| : c_1, c_2 \in \mathcal{C}\}$. The set of available transmit power levels depends on the hardware used at the AP. In some situations, the set of power levels can be further restricted to a subset by the network administrator. The set of available channels is primarily defined by the standard. The IEEE 802.11g standard defines a spectrum with 13 channels. However, channel availability may vary in accordance with country regulations and the local environment. Although, the set of possible transmit power levels and the set of available channels may vary by AP, to simplify the presentation we assume that the sets are the same for all APs.

The service area is modeled by a set of grid points that represent a set of locations where the network coverage is desired. The grid points are therefore referred to as coverage test points (TPs). The set of TPs is denoted by $\mathcal{J}$. The attenuation between a candidate AP location a and each TP is given in a path-loss prediction grid and is represented as a set of power gain values $\{g_{aj}, j \in \mathcal{J}\}$.

Next, for each AP and every given transmit power level we define two ranges. A *serving range* of an AP is an area where the received signal from the AP is sufficiently strong to get an MT associated to the AP. The serving range of AP a is denoted by $\mathcal{R}_{al}^{srv}$, where l is the transmit power of a, and is defined by the serving threshold γ^{srv}. (Note that although l is a power level, i.e., a real number, we use it also as a subscript.) The threshold defines the minimal signal strength required for receiving transmissions at the lowest desired data rate.

The threshold is usually an adjustable configuration parameter and its typical values for a specific hardware can be found in hardware specifications, e.g., in [16].

The second range is used to define an area where STAs (either MTs or other APs) can hear the AP and detect the medium as busy in case the AP transmits on the same channel. The range is related to physical carrier sensing and is therefore called *CS range*. We use $\mathcal{R}_{al}^{cs}$ to denote the CS range of AP a transmitting at power level l. We assume that physical carrier sensing is based on detecting other STAs' DSSS signals (CCA Mode 2). The CS range is defined by the CS threshold γ^{cs}. For simplicity, we assume that γ^{srv} and γ^{cs} do not vary by station. Mathematically, the two signal ranges of AP a transmitting at power level l are defined as follows,

- serving range: $\mathcal{R}_{al}^{srv} = \{j \in \mathcal{J} : l \cdot g_{aj} \geq \gamma^{srv}\}$,
- CS range: $\mathcal{R}_{al}^{cs} = \{j \in \mathcal{J} : l \cdot g_{aj} \geq \gamma^{cs}\}$.

With the two ranges defined above, we assume that an MT at TP j can be served by AP a if and only if j is within the AP's serving range, i.e., $j \in \mathcal{R}_{al}^{srv}$, where l is the transmit power level of AP a. If the received signal strength is below γ^{srv} but is at least γ^{cs}, the carrier is still sensed as busy when AP a transmits, restraining the MT from transmitting. That is, $\mathcal{R}_{al}^{srv} \subset \mathcal{R}_{al}^{cs}$.

User throughput in WLANs depends on a number of factors among which the most important are network congestion, contention from other STAs, and interference. Most of the factors are dynamic which makes it difficult to design a static plan of a WLAN. In this situation, the coverage measure and the net throughput are very helpful. When channel assignment is to be decided, coverage overlap is often used as an interference measure (see, for example, [55]). Below we define coverage, net throughput, and overlap.

Coverage. We consider that TP j ($j \in \mathcal{J}$) is covered if there is installed at least one AP a ($a \in \mathcal{A}'$) such that the received signal from a is at least at a level of the serving threshold γ^{rsv}. Thus, depending on network topology and configuration, a TP is *covered* if it belongs to the serving range of at least one installed AP, and *not covered* otherwise. A network with (almost) complete coverage is desirable, i.e., more than 95–97 % of the area is typically to be covered.

Net throughput. To model net throughput, we use our measurements to find a polynomial fitting function (see Section 7.3) that represents the throughput experienced by a user in a contention-free environment. Let $\phi(\cdot)$ denote the function that models net user throughput. The throughput depends on the strongest signal received from covering APs. For an MT, a serving AP is the AP from which the strongest signal is received. The net throughput experienced by a user at TP j is thus given by $\phi(l, g_{aj})$, where a is the serving AP, and l is its transmit power. Low net throughput in some area is a strong indication of that a better AP placement is needed. Therefore, one of our objectives amounts to maximizing the total net throughput over all TPs by choosing an appropriate subset of candidate AP locations.

Overlap. Channel assignment is our second design goal. In WLANs, the channels need to be assigned in a way such that STAs interfere with each other as little as possible and their contention for the medium in the network is minimized. This implicitly improves the network throughput. To reduce contention probability, one of the most common approaches is to minimize the overlap of the coverage areas of APs operating on the same channel. For APs operating on adjacent channels, the overlap may be also taken into account to reduce adjacent channel interference. Although in some situations reducing interference may be more vital, minimizing contention is typically more addressed.

To model coverage overlap, we use the CS threshold γ^{cs} defining the minimum received signal power for sensing the channel as "busy". In this chapter, AP transmit power is fixed

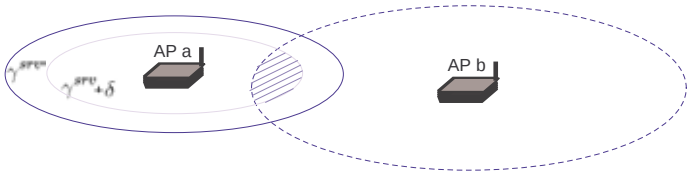


Figure 7.1: Coverage overlap area of two APs.

and is the same for all APs (let it be denoted by L). The actual number of overlapping APs at a given TP depends on the AP placement and the channel assignment in the network. For optimization, the overlap is computed for each pair of APs and estimated in number of TPs. The overlap of two APs, a and b, is found as follows,

$$\nu_{ab} = \left|\left\{j \in \mathcal{R}^{cs}_{aL} \cap \mathcal{R}^{cs}_{bL} : \max\{L \cdot g_{aj}, L \cdot g_{bj}\} \geq \gamma^{srv} + \delta)\right\}\right|. \tag{7.1}$$

By (7.1), the overlap between APs a and b is a parameter that equals the number of TPs at which both APs are detectible (i.e., the TPs must be within the CS range of each of the two APs) and the signal from at least one of the APs is above serving threshold γ^{srv}. A margin denoted by δ is added to the threshold for two reasons. First, to reduce overlap overestimation, i.e., we do not count TPs if none of the two APs is likely to be the serving AP. Second, to overcome dynamic variations of the signal strength. The recommended range for the δ-parameter is 10–20 dB. Figure 7.1 shows the overlap area of APs a and b (the overlap is represented by the shaded area).

7.2 Optimization Problems

In this section, we start with optimization problems addressing the AP placement task and the channel assignment task separately and then present how the models can be combined into a more complex optimization model that allows us to find an optimal network plan addressing the two network planning tasks simultaneously.

7.2.1 AP Placement

Two models are presented in this section. Both models are used to decide on AP locations, i.e., to choose a subset from a set of given candidate AP locations. By the first model (denoted by W1-NAP) the number of installed APs is to be minimized subject to a coverage constraint. In the second model, the average net user throughput is to be maximized such that the number of installed APs does not exceed M. Since we do not consider contention or interference in the two models, the maximum transmit power of APs is assumed.

Minimum Number of Installed APs

From the optimization stand point, pure coverage problems are viewed as set-covering problems. The classical set covering problem has been discussed in Section 1.2.2. In this section, the problem is used to find the minimum amount of equipment needed to provide full coverage in the planned service area. Although, the set covering problems are known to be $\mathcal{NP}$-hard [42], models of this type can be efficiently solved for very large instances using integer programming techniques [14].

To formulate the problem mathematically, we use the following set of variables.

$$z_a = \begin{cases} 1 & \text{if AP } a \text{ is installed,} \\ 0 & \text{otherwise.} \end{cases}$$

The problem formulation is then as follows.

$$\text{[W1-NAP]} \qquad \sum_{a\in\mathcal{A}'} z_a \longrightarrow \min \tag{7.2a}$$

$$\text{s.t.} \qquad \sum_{\substack{a\in\mathcal{A}':\\ j\in\mathcal{R}_{aL}^{srv}}} z_a \geq 1 \qquad j\in\mathcal{J} \tag{7.2b}$$

$$z_a \in \{0,1\} \qquad a\in\mathcal{A}' \tag{7.2c}$$

In W1-NAP, the objective (7.2a) is the number of installed APs. Constraints (7.2b) ensure full coverage of the service area, i.e., each TP is covered with a sufficiently strong signal from at least one AP, and (7.2c) are the integrality constraints.

Maximum Net Throughput

As defined in Section 7.1, net throughput is the maximum achievable throughput a user can expect in ideal conditions, i.e., when there is no other STA contending for the medium. Net throughput depends on the transmission data rate which, it turn, depends on the distance between the serving AP and the user (a more detailed study on net throughput is presented in Section 7.3). The AP placement thus determines the maximum user throughput in different parts of the network. Our optimization goal here is to select AP locations of at most M APs such that the average net user throughput taken over all TPs is maximized. The problem can be viewed as the maximum k-facility location problem where at most k facilities are to be installed such that the total profit is maximized (see Section 1.2.2 for more details on facility location problems).

To formulate the problem, in addition to the z-variables used in model W1-NAP, we introduce one more set of variables defined as follows.

$$x_{aj} = \begin{cases} 1 & \text{if TP } j \text{ is served by AP } a,\\ 0 & \text{otherwise.}\end{cases}$$

The complete maximum net throughput model is formulated below.

$$\text{[W1-NT]} \qquad \frac{1}{|\mathcal{J}|}\cdot\sum_{a\in\mathcal{A}'}\sum_{j\in\mathcal{R}_{aL}^{srv}} \phi(L,g_{aj})x_{aj} \longrightarrow \max \tag{7.3a}$$

$$\text{s.t.} \qquad x_{aj} \leq z_a \qquad \forall\, a\in\mathcal{A}', j\in\mathcal{R}_{aL}^{srv} \tag{7.3b}$$

$$\sum_{\substack{a\in\mathcal{A}':\\ j\in\mathcal{R}_{aL}^{srv}}} x_{aj} \leq 1 \qquad \forall\, j\in\mathcal{J} \tag{7.3c}$$

$$\sum_{a\in\mathcal{A}'} z_a = M \tag{7.3d}$$

$$z_a \in \{0,1\} \qquad \forall\, a\in\mathcal{A}' \tag{7.3e}$$

$$x_{aj} \in \{0,1\} \qquad \forall\, a\in\mathcal{A}', j\in\mathcal{R}_{aL}^{srv} \tag{7.3f}$$

In W1-NT, the objective function (7.3a) measures the average net throughput per TP. Constraints (7.3b) ensure that a TP cannot be served by an AP which is not installed. By constraints (7.3c), a TP can be served by at most one AP. Exactly M APs are to be installed by constraint (7.3d). Constraints (7.3e) and (7.3f) are the integrality constraints.

The problem is $\mathcal{NP}$-hard which can be shown by reducing the $\mathcal{NP}$-complete node covering problem (see, for example, [43]). For WLAN instances of the sizes we are interested in, the optimization problem can quickly be solved to optimality using a standard MIP solver. In practice, however, it may be interesting to study larger networks. In this case, using state-of-the-art facility location algorithms and refined models is encouraged.

7.2.2 Channel Assignment

Once the AP locations are decided, the next natural step for the network planner is to configure the network. Channel assignment is one of the major tasks in this stage. One very common way of modeling the channel assignment is to view it as a graph coloring problem, e.g., [8, 23, 49, 57]. In a simplified approach, only non-overlapping channels are assumed to be available at each AP. As depicted in Figure 6.3, these are channels 1, 6, and 11 in 802.11b/g WLANs. In any such channel assignment, the three channels are interchangeable, and only assigning the same channel to APs with the overlapping CS ranges may result in a negative effect on the objective. More sophisticated models take also into account interference on adjacent channels, which, first of all, being less restrictive, gives the network planner more flexibility in deciding on channel assignment, and second, may significantly improve the network performance. In a plane or one-floor scenario, co-channel overlap is more vital in WLANs than in GSM networks, for example, where the number of available channels is much larger and co-channel interference can thus be reduced to a small amount. However, in a multi-floor scenario (similar to or even larger than the one considered in this chapter), minimizing adjacent channel interference becomes as important as co-channel overlap reduction. Next, we present a channel assignment model that addresses both aspects and can be used for any channel set that may also vary by AP.

Minimum Weighted Channel Overlap

To formulate the problem, we use (binary) decision variables $\{f_a^c, a \in \mathcal{A}, c \in \mathcal{C}\}$ to define which channel is to assigned to which AP, i.e., the set of variables is defined as follows,

$$f_a^c = \begin{cases} 1 & \text{if AP } a \text{ operates on channel } c, \\ 0 & \text{otherwise.} \end{cases}$$

To take overlap on adjacent channels into account, we introduce a binary variable w_{ab}^d for each pair of APs (a, b) that may overlap (i.e., $\mathcal{R}_{aL}^{cs} \cap \mathcal{R}_{bL}^{cs} \neq \emptyset$) and each channel distance $d \in \mathcal{D}$. The variables are defined in the following way,

$$w_{ab}^d = \begin{cases} 1 & \text{if the channel distance for AP } a \text{ and AP } b \text{ equals } d, \\ 0 & \text{otherwise.} \end{cases}$$

To avoid redundancy in the formulation, the set of the w-variables is defined for ordered pairs of APs. The pairs are the elements of the following set,

$$\mathcal{A}^2 = \{(a, b) : a, b \in \mathcal{A}, a < b\} .$$

The complete model, further referred to as the minimum weighted channel overlap model and denoted by W1-WCO, is formulated as follows.

$$[\text{W1-WCO}] \quad \sum_{(a,b)\in\mathcal{A}^2} \sum_{d\in\mathcal{D}} \nu_{ab} F(d)\, w_{ab}^d \longrightarrow \min \tag{7.4a}$$

$$\text{s. t.} \quad \sum_{c\in\mathcal{C}} f_a^c = 1 \qquad \forall\, a \in \mathcal{A} \tag{7.4b}$$

$$f_a^{c_1} + f_b^{c_2} \leq 1 + w_{ab}^{|c_1 - c_2|} \qquad \forall\, (a, b) \in \mathcal{A}^2, c_1, c_2 \in \mathcal{C}, |c_1 - c_2| \in \mathcal{D} \tag{7.4c}$$

$$w_{ab}^d \in \{0, 1\} \qquad \forall\, (a, b) \in \mathcal{A}^2, d \in \mathcal{D} \tag{7.4d}$$

$$f_a^c \in \{0, 1\} \qquad \forall\, a \in \mathcal{A}, c \in \mathcal{C} \tag{7.4e}$$

In W1-WCO, constraints (7.4b) ensure that a single channel is chosen for each AP, and constraints (7.4c) set the relation between channels assigned to two APs and the corresponding w-variables. Integrality of the w- and f-variables is stated by constraints (7.4d) and (7.4e),

respectively. The objective function is the sum of weighted overlap metrics for all pairs of APs. The overlap metrics ν_{ab} are defined by (7.1). To allow for assigning overlapping channels, the weighting coefficients in the objective function are modeled by a monotonically decreasing function $F(d)$ defined such as $F(0) = 1$ and $0 < F(d) < 1$, for any channel distance $d > 0$. Modeled in this way, function $F(d)$ allows us to prioritize a larger channel distance for neighboring APs. In our computational experiments we used $F(d) = 1/(1+d)^k$, where k ($k > 0$) is a parameter that controls the impact of assigning overlapping channels to neighboring APs. To focus on the impact of small channel distances, a $k > 1$ should be chosen. (We use $k = 2$ in our numerical experiments.) Alternatively, function $F(d)$ could also be defined empirically (see, for example, [28, 55]).

Observe that W1-WCO is a problem from the class of minimum-interference frequency assignment problems studied, for example, in [8, 23, 21, 46]. As a special case, the model can also be used for a set of non-overlapping channels by restricting set $\mathcal{D}$ to a single element, i.e., channel distance of zero. For this special case, the problem can also be formulated without f-variables by adopting, for example, formulation of the *minimum k-partition* problem as it has been done in [22]. The formulation for three non-overlapping channels is presented below and is further referred to as the minimum co-channel overlap problem.

Minimum Co-Channel Overlap

The following set of decision variables is used in the formulation of the minimum co-channel overlap problem,

$$w_{ab} = \begin{cases} 1 & \text{if AP } a \text{ and AP } b \text{ operate on the same channel,} \\ 0 & \text{otherwise.} \end{cases}$$

Note that only one w-variable is needed for each two APs a and b. Therefore, the w-variables are defined for set $\mathcal{A}^2$ of ordered pairs.

The complete formulation reads as:

$$\text{[W1-CO]} \qquad \sum_{(a,b)\in\mathcal{A}^2} \nu_{ab} w_{ab} \longrightarrow \min \tag{7.5a}$$

$$\text{s.t.} \qquad \sum_{\substack{\{a,b\}\subset Q:\\ (a,b)\in\mathcal{A}^2}} w_{ab} \geq 1 \qquad \forall\, Q \subseteq \mathcal{A}, |Q| = 4 \tag{7.5b}$$

$$w_{ab} + w_{bc} \leq 1 + w_{ac} \qquad \forall\, (a,b),(b,c) \in \mathcal{A}^2 \tag{7.5c}$$

$$w_{ab} + w_{ac} \leq 1 + w_{bc} \qquad \forall\, (a,b),(b,c) \in \mathcal{A}^2 \tag{7.5d}$$

$$w_{ac} + w_{bc} \leq 1 + w_{ab} \qquad \forall\, (a,b),(b,c) \in \mathcal{A}^2 \tag{7.5e}$$

$$w_{ab} \in \{0,1\} \qquad \forall\, (a,b) \in \mathcal{A}^2 \tag{7.5f}$$

The objective (7.5a) minimizes the overlap area of APs operating on the same channel. As in W1-WCO, the overlap metrics ν_{ab} are defined by (7.1). By (7.5b), among any four APs, at least two APs are assigned the same channel. Constraints (7.5c)-(7.5e) are triangle inequalities ensuring that for any three APs, two pairs of APs cannot be assigned the same channel while the third pair is assigned a different channel. Constraints (7.5f) ensure integrality of w-variables. Given a solution to W1-CO, assigning channels is done based on the observation that the graph formed by non-zero w-variables in the solution consists of at most three cliques [22]. Each clique is associated with one channel, and all vertices in the clique are assigned this channel.

The disadvantage of W1-CO over W1-WCO is the exponential growth of the number of constraints (7.5b) with the number of non-overlapping channels $|\mathcal{C}|$ while $|\mathcal{C}| \leq \frac{|\mathcal{A}|}{2}$ (note that $|Q| = |\mathcal{C}| + 1$). Also, to ensure feasible channel assignment, the problem is to be solved for a complete graph which is inefficient for sparse networks. On the other hand, the great

advantage is that the formulation eliminates symmetry which is present in W1-WCO because of the f-variables.

7.2.3 Integrated AP Placement and Channel Assignment

The models presented in Sections 7.2.1 and 7.2.2 separately address optimizing AP placement and channel assignment. In this section, both aspects are taken into account together by combining models of the two types into a single model. Simultaneous optimization for the two aspects is expected to give a better network design solution. Mathematically, the new objective function is formulated as linear combination of the two components associated with the two network design tasks, i.e., selecting AP locations and assigning the channels. In case two APs transmit on the same frequency, their coverage overlap is deducted from the throughput gain. A trade-off parameter $\alpha \in (0,1)$ specifies the relative weights of the two optimization goals. (Note that there is no need to solve the integrated model for the two extreme cases, i.e., when $\alpha = 0$ or $\alpha = 1$, since the problem reduces to W1-NT and W1-WCO, respectively.) Observe that the trade-off parameter gives an adequate measure of relative importance of the two goals only if the absolute values of the two combined objectives are approximately the same. Therefore, some additional scaling parameter K is introduced. The value of K is dependent on the data instance and is to be found empirically.

Combining models W1-NT and W1-WCO, the integrated model (denoted by W1-NTWCO) maximizes net user throughput by deciding AP locations and optimizes channel assignment by minimizing the sum of penalties from assigning the same or adjacent channels to APs with overlapping CS ranges. The combined model interacts only in the objective function and in one type of constraints. For the sake of convenience, we define a set of ordered pairs of candidate AP locations (similar to $\mathcal{A}^2$) as follows,

$$\mathcal{A}'^2 = \{(a,b) : a,b \in \mathcal{A}', a < b\} \ .$$

The integrated problem formulation is presented below.

$$\text{[W1-NTWCO]} \quad (1-\alpha)K \sum_{a\in\mathcal{A}'} \sum_{j\in\mathcal{R}_{aL}^{srv}} \phi(L, g_{aj}) x_{aj} - \alpha \sum_{(a,b)\in\mathcal{A}'^2} \sum_{d\in\mathcal{D}} \nu_{ab} F(d)\ w_{ab}^{d} \longrightarrow \max \tag{7.6a}$$

$$\text{s.t.} \quad x_{aj} \le z_a \qquad \forall\ a \in \mathcal{A}', j \in \mathcal{R}_{aL}^{srv} \tag{7.6b}$$

$$\sum_{\substack{a\in\mathcal{A}':\\ j\in\mathcal{R}_{aL}^{srv}}} x_{aj} \le 1 \qquad \forall\ j \in \mathcal{J} \tag{7.6c}$$

$$\sum_{a\in\mathcal{A}'} z_a = M \tag{7.6d}$$

$$f_a^{c_1} + f_b^{c_2} \le 1 + w_{ab}^{|c_1-c_2|} \qquad \forall\ (a,b) \in \mathcal{A}'^2, c_1, c_2 \in \mathcal{C}, |c_1 - c_2| \in \mathcal{D} \tag{7.6e}$$

$$\sum_{c\in\mathcal{C}} f_a^c = z_a \qquad \forall\ a \in \mathcal{A}' \tag{7.6f}$$

$$z_a \in \{0,1\} \qquad \forall\ a \in \mathcal{A}' \tag{7.6g}$$

$$x_{aj} \in \{0,1\} \qquad \forall\ a \in \mathcal{A}', j \in \mathcal{R}_{aL}^{srv} \tag{7.6h}$$

$$w_{ab}^{d} \in \{0,1\} \qquad \forall\ (a,b) \in \mathcal{A}'^2, d \in \mathcal{D} \tag{7.6i}$$

$$f_a^c \in \{0,1\} \qquad \forall\ a \in \mathcal{A}', c \in \mathcal{C} \tag{7.6j}$$

In W1-NTWCO, objective (7.6a) combines objectives (7.3a) and (7.4a). AP selection is linked to channel assignment in (7.6f), by which an AP is to be assigned a channel only if being installed.

Combining the net throughput model with the second channel assignment model described in Section 7.2.2 is straightforward [22]. Similar to W1-NTWCO, the two models are linked in

the new objective function and in one type of the constraints. In particular, constraints (7.5b) change to

$$\sum_{\substack{\{a,b\}\subset Q:\\ (a,b)\in\mathcal{A}'^2}} w_{ab} \geq \sum_{a\in\mathcal{Q}} z_a - 3 \qquad \forall\, Q \subseteq \mathcal{A}', |Q| = 4\ .$$

7.3 An Experimental Study on User Net Throughput

7.3.1 Network Configuration

A study on the user net throughput in an IEEE 802.11g WLAN operating in the 2.4 GHz frequency band was conducted within a COST Action TIST 293 STSM project [61]. The studied network is a WLAN deployed at Zuse Institute of Berlin (ZIB). The WLAN consists of 32 APs connected to an Ethernet (100 Mbps): 29 APs are along the main floors (seven/eight APs per floor), one AP is at the roof floor, and two APs are at the underground floor. 19 out of 32 APs are of type Cisco AP-1200/AP21G [16] and support both IEEE 802.11b and IEEE 802.11g standards. Four of them use directional antennas that act as remote radio interfaces. The rest are of type Cisco AP-350 [17] and support only IEEE 802.11b.

In the current network configuration, the APs are placed in an ad-hoc manner following a set of simple rules. The APs are uniformly distributed along corridors in each floor, and are placed mainly in service rooms in the center part of the corridors. To avoid user association to APs from an adjacent floor (assuming that such connections are most likely of low quality), the APs are placed one under another among the floors. Therefore, in the resulting network configuration the BSSs are separated by floors. A drawback of this approach is that the neighboring APs in the vertical dimension become strong interferers to each other (if assigned different channels).

In the most part of the building, coverage by at least one AP is achieved, but in many places strong signals from several APs, sometimes on the same channel as the serving AP, were present. If the signals are strong enough, this may cause a ping-pong effect, i.e., that an MT frequently switches between the covering APs. If the strongest received signal is not sufficiently strong to overcome the interference, the wireless connection will be unstable. The main problem we observed in the current network configuration is that the wireless connection is significantly better (in general) in corridors than in offices, although the latter should be of a higher importance. This is a strong indication of that a better AP placement is needed.

7.3.2 Measurement Experiments Setup

To study the relation between the received signal strength and the net user throughput, we experimented with TCP transmissions between different APs (IEEE 802.11g) and an MT. Five sets of experiments with six runs in each set were designed. The range [-90, -20] dBm is divided into five equal intervals. Each interval corresponds to one set of experiments such that the strongest received signal in the experiments falls into the interval. For the signal measurements we used Cisco Aironet Desktop Utility (ADU) installed on a laptop equipped with a Cisco Aironet 802.11a/b/g Wireless LAN Client adapter [18].

At each selected location, we performed six measurement runs in a row using the network benchmarking tool Netio [59]. In each run, we measured TCP throughput by continuously transmitting TCP packets of 1 KB during one minute between the laptop and a PC connected to a Gigabit Ethernet network. Since the tested APs (Cisco AP-1200/AP21G [16]) were also connected to the fixed network, the radio link between the AP and the laptop was the bottleneck. The experiments were performed in absence of contending MTs, which allows us to treat the obtained throughput measurements as user net throughput. The measurements were done for both DL and UL. Since we did not observe any significant difference between throughput in the two directions, we further focus on the DL direction.

7.3.3 Results

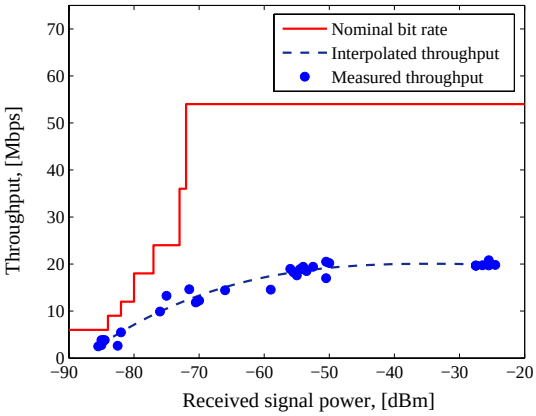


Figure 7.2: Nominal bit rate and net throughput vs. received signal power.

Figure 7.2 shows the measured net throughput and nominal IEEE 802.11g bit rates [16] versus the received signal strength. In spite of no contention during the experiments, the highest throughput (achieved when the laptop was placed just in front of the AP) is around 20 Mbps, which is less than half of the highest 802.11g bit rate (54 Mbps). Such a low net throughput is mainly explained by transmission delays and protocol overhead caused by the CSMA-CA RTS/CTS mechanism. Transport layer protocol (TCP in our experiments) also affects the net throughput. Our observations are compliant with conclusions in [30] and [41].

Further, we find a least-squares fit to the set of measured throughput values. The fitting curve is represented by the following third-degree polynomial,

$$\varphi(\rho) = 0.0001 \cdot \rho^3 + 0.0069 \cdot \rho^2 + 0.1586 \cdot \rho + 20.9542 \ , \tag{7.7}$$

where $\varphi(\rho)$ is the net user throughput in DL (in Mbps) as a function of the received signal power ρ (in dBm). Note that the relation between $\varphi(\rho)$ and the net user throughput parameter used in models W1-NT and W1-NTWCO is as follows,

$$\phi(L, g_{aj}) = \varphi(10 \log_{10}(L \cdot g_{aj}) + 30) \ . \tag{7.8}$$

7.4 Numerical Experiments

In this section, we experiment with the optimization models presented in Section 7.2. All numerical experiments were conducted on a standard PC (2.4 GHz Pentium processor, 2 GB RAM). The optimization models were compiled using ZIMPL [44] as the modeling language and solved to optimality with ILOG CPLEX 10.0 [63]. A visualization tool kindly provided by Hans-Florian Geerdes, Zuse Institute of Berlin, was used to generate Figures 7.3, 7.4, and 7.6-7.9.

The test network is described in Section 7.4.1. In the first part of our study, we find the minimum number of APs necessary to ensure full coverage depending on the required minimum received signal strength. Next, we compare a solution obtained by the two-step network planning approach to that obtained by joint optimization of AP locations and channel assignment. Furthermore, we experiment with different values of the trade-off parameter α in the combined model and study their effect on the solutions.

7.4.1 Test Network and a Reference Scenario

Our test network represents a part of the WLAN deployed at Zuse Institute of Berlin (see Section 7.3.1 for more details on the network). There are 32 candidate AP locations for this two-floor scenario as shown in Figure 7.3. We assume that all APs are of type Cisco AP-1200/AP21G [16] and compliant with the IEEE 802.11g standard. Every AP is equipped with an omni-directional antenna. The path-loss predictions were obtained for each candidate AP location via 3D ray-tracing methods with multiple reflections using a 3D model of the building [37, 38]. An example of a path-loss prediction is demonstrated in Figure 7.3. Table 7.1 summarizes network statistics.

As a reference scenario, we consider a network configuration similar to that used in the corresponding part of the real network. In this scenario, there are eight APs installed. Three non-overlapping channels (1, 6, and 11) are used as the base in the channel assignment, although channel 7 is additionally used to reduce contention in the network. The estimated user throughput averaged over the studied area is 10.69 Mbps. Due to the AP placement strategy, by which APs are installed only in the center part of corridors in service rooms with thick concrete walls, most parts of the area experience low throughput. In particular, net user throughput is below 1 Mbps in 22.75 % of the entire area, among which 11.5 % is due to coverage loss, i.e., the received signal drop below γ^{srv}. To evaluate a network design, we also compute contention probability at each TP as the relative size of the entire area from where contention may be expected. In the reference scenario, the average contention probability over all covered TPs is 0.18.

The reference scenario is depicted in Figures 7.7(a), 7.7(b), 7.7(c), and 7.7(d), which demonstrate the best server map, net throughput distribution, overlap graph, and contention map, respectively. The information visualized in the figures is described below.

- *Best server map.* Best server map is created for a set of installed APs. The color of each pixel in the best server map corresponds to the channel assigned to the AP from which the received signal is the strongest in the pixel. From the best server map, it is easy to see the coverage area of each AP.
- *Net throughput distribution.* Net throughput is computed by (7.7) for each pixel with respect to its best server.
- *Overlap graph.* Each node in the overlap graph represents an AP, and color intensity of edges depicts the amount of overlap between two APs, i.e., represents a value of the corresponding overlap parameter ν_{ab}. If two APs do not overlap, there is no edge between

Table 7.1: Test network characteristics

Characteristic / Parameter	Value
Number of floors	2
Service area size in each floor [m×m]	84×18
Number of candidate locations $\lvert\mathcal{A}'\rvert$	32
Number of installed APs M	8
AP height above the floor [m]	2
AP antenna type	AIR-ANT-4941
AP antenna gain [dBi]	2
AP transmit power [mW]	30
Frequency band [GHz]	2.4
Channel set	13 channels (ETSI), 2.412–2.484 GHz
Number of TPs	798
TP grid resolution [m]	2
TPs' height above the floor [m]	1
Serving threshold [dBm]	−90
CS threshold [dBm]	−115

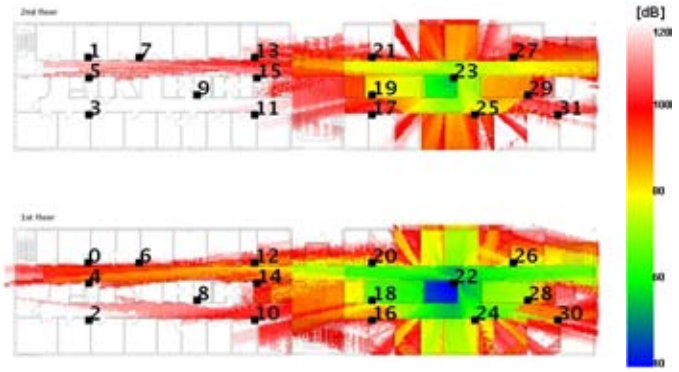


Figure 7.3: Candidate APs locations and path-loss predictions for location 22.

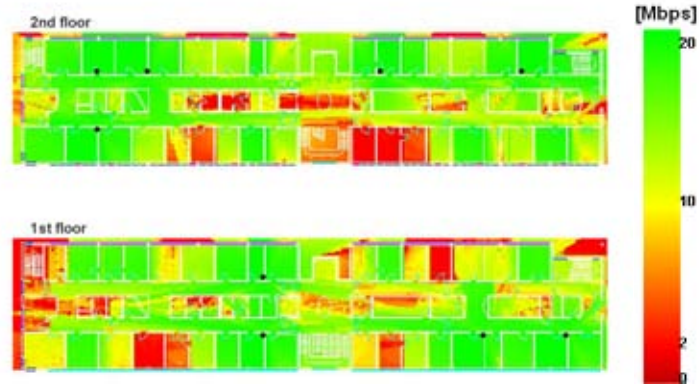


Figure 7.4: User net throughput for a network design with 9 APs subject to the full coverage requirement with $\gamma^{srv} = -90\,\text{dBm}$ (average: 13.23 Mbps; areas with net throughput below 1 Mbps: 6.05 %; coverage loss: 0 %.)

the corresponding nodes in the graph. If the overlap is too small, the corresponding edge is almost fully transparent. Thus, it is desirable to have nearly transparent edges between APs assigned the same channel or having a small channel distance, whilst large overlap is typically not harmful when the channel distance is large enough.

- *Contention map.* The color of each pixel represents the size of the area from where contention can be expected. Only contention through APs is considered: the size of the coverage area of an AP adds to the contention measure for a TP, if the TP is assigned the same channel as the AP and the TP is within the AP's CS range. White color in the scale corresponds to the average AP coverage area computed as $|\mathcal{J}|/|\mathcal{A}|$.

As previously mentioned, the number of installed APs in the reference scenario is eight. Therefore, when optimizing network configuration (see Sections 7.4.2 and 7.4.3), we also require the total number of installed APs to be at most $M = 8$. In this section, however, we study how the minimum number of APs is affected by the serving threshold.

We evaluate the minimum amount of APs needed to provide complete coverage of the given area for a given serving threshold γ^{srv}. For this, we use optimization model W1-

NAP presented in Section 7.2.1. The thresholds were selected to represent the sensitivity thresholds for certain bit rates taken from Cisco documentation [16] to ensure the minimum best-case transmission bit rate which we use as a coverage quality measure. The obtained optimal solutions are presented in Table 7.2.

The last column of the table shows CPU times. We observe that computational times significantly vary by threshold. This is explained by the amount of preprocessing reduction made by Cplex. The reduction is larger for higher threshold values since the number of possible servers for each TP decreases. Moreover, when increasing the serving threshold, some bins can never be covered with the given set of candidate AP locations. These bins were excluded from the set of TPs to enable feasible solutions to the problem. The number of such bins is shown in the fourth column of the table.

From the results, we observe that at least nine APs are required to achieve a bit rate of at least 6 Mbps at any TP, i.e., to ensure non-zero user net throughput in the network (according to Figure 7.2). In the real WLAN, the number of installed APs in the same area is eight, which therefore may result in no wireless access or unstable wireless connections in some parts of the area. Note, however, that the net throughput achieved by a user at the best is significantly lower than the transmission bit rate, and that the real throughput is even lower if the user's MT experiences high interference and/or has to contend to access the medium.

The net throughput distribution over the service area for the first solution in Table 7.2 (i.e., when $\gamma^{srv} = -90$ dBm) is presented in Figure 7.4 where installed APs are marked by black circles. In this solution, the average net throughput over the area is 13.23 Mbps. The areas with potentially low user throughput (net throughput below 1 Mbps) sum up to 6.05 % of the total area, which is a significant improvement over the reference scenario (compare to Figure 7.7(b)).

Further, we intend to show that even with the same number of installed APs as in the reference scenario (i.e., assuming $M = 8$), the network performance still can be significantly improved, although with no full coverage requirement.

7.4.2 Two-step Optimization

In this section, we follow a traditional multi-step network design approach, and optimize the network configuration in two steps. In particular, we first solve the AP placement problem and then optimize the channel assignment. For the first step, we use model W1-NT that has been presented in Section 7.2.1, i.e., we find M AP locations such that user net throughput averaged over the service area is maximized. A model with three non-overlapping channels, i.e., a typical scenario in practice, is considered for channel assignment optimization in the second step. This is a special case of the minimum weighted channel overlap model, W1-WCO, with $\mathcal{C} = \{1, 6, 11\}$ and $\mathcal{D} = \{0\}$.

Table 7.2: Optimal solutions to W1-NAP for different serving thresholds

Serving threshold γ^{srv}, [dBm]	Bit rate [Mbps]	Number of installed APs	Number of excluded TPs	CPU time [sec]
-90	6	9	0	3.28
-84	9	12	1	0.68
-82	12	13	1	0.42
-80	18	16	2	0.13
-77	24	18	4	0.09
-73	36	20	6	0.06
-72	48 or 54	21	8	0.05

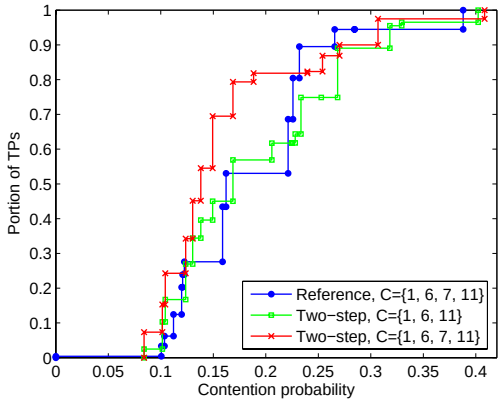


Figure 7.5: Contention distribution in the reference scenario and two two-step solutions.

Figures 7.8(a)-7.8(d) demonstrate a network design found by the two-step approach. Comparing Figures 7.8(a) and 7.7(a), we observe that, unlike in the reference scenario where the service area of each AP is mostly limited to the floor where the AP is installed, all APs in the optimized network design serve some areas on both floors.

The total coverage loss in the obtained solution is 1.71 %. However, if we compare the net throughput statistics for this solution to the corresponding statistics of the solution with 9 APs obtained in Section 7.4.1, this is a moderate price for maintaining net throughput at almost the same level (13.16 Mbps vs 13.23 Mbps) with fewer APs. Furthermore, the total size of areas with low net throughput does not significantly increase (6.48 % vs 6.05 %; note that 6.48 % also includes coverage loss of 1.71 %). From the interference and contention points of view, the solution with 8 APs obtained by sequential optimization is better than that with 9 APs obtained in Section 7.4.1, provided that the same channel set, i.e., $\mathcal{C} = \{1, 6, 11\}$, is used in both solutions.

Compared to the reference scenario, we observe a significant improvement in the user net throughput distribution (see Figure 7.7(b)). In particular, the average net throughput over the area is 23.11 % higher in the obtained solution than that in the reference scenario (compare 13.16 Mbps to 10.69 Mbps). Moreover, two-step optimization reduces the total size of areas that either are not covered or suffer from low net throughput (below 1 Mbps) by 3.5 (compare 22.75 % to 6.48 %).

Considering contention (in the way it has been defined in Section 7.4.1), we observe that contention distribution in Figure 7.8(d) is slightly worse than that in the reference scenario (see Figure 7.7(d)), which is mainly explained by a smaller number of channels in the optimized solution (three channels vs four channels in the reference scenario). The average contention probability in the optimized three-channel assignment is 0.18.

For a proper comparison to the reference scenario, we next solve the channel assignment problem for four channels used in the reference scenario (i.e., with $\mathcal{C} = \{1, 6, 7, 11\}$ and $\mathcal{D} = \{0, 1\}$) for the same AP locations found by W1-NT. The obtained channel assignment is shown in Figure 7.8(c) (in red color). The average contention probability is improved by 11 % as compared to the reference scenario (0.16 vs 0.18). Figure 7.5 demonstrates cumulative distributions of contention probability in the reference scenario, three-channel solution, and four-channel solution.

The computational time for solving the AP location problem (by W1-NT) did not exceed 30 s. For the channel assignment problem (solved with W1-WCO), the computational time was less than one second, even for four channels. The optimal solutions obtained by

W1-CO are the same as those obtained by W1-WCO with three non-overlapping channels, although the computational time was slightly faster, as expected.

7.4.3 Joint Optimization

Our studies show that performance of the reference configuration can further be improved by jointly optimizing AP locations and channel assignment. This is done by integrated model W1-NTWCO presented in Section 7.2.3. Furthermore, we investigate the influence of each of the optimization objectives on key performance metrics by varying their relative weights controlled by the trade-off parameter.

In our numerical experiments, we let α vary in the interval $[0, 1]$ with step of 0.1. Small values of α emphasize throughput maximization, while values close to 1.0 stress overlap minimization. The two extreme cases are handled separately. Thus, we solve W1-NT and W1-WCO sequentially when $\alpha = 0$ (as has been discussed in Section 7.4.2), and we solve W1-NTWCO but then manually assign the x-variables when $\alpha = 1$. All obtained solutions are also compared to the reference scenario. The results for three non-overlapping channels (i.e., $\mathcal{C} = \{1, 6, 11\}, \mathcal{D} = \{0\}$) and for three channels assuming that they do overlap (i.e., $\mathcal{C} = \{1, 6, 11\}, \mathcal{D} = \{0, 5, 10\}$) are presented in Table 7.3 and Table 7.4, respectively. In some intervals, different values of α lead to the same solution, so the corresponding columns are grouped. For each presented solution, we compute the two objective functions, i.e., W1-NT and W1-WCO, and the following performance statistics:

- *Coverage loss* - the fraction of the area where the strongest received signal drops below the serving threshold (the rest of the area is thus considered covered);
- *Average net throughput* - net user throughput averaged over the covered area (note the difference with W1-NT which provides the average net user throughput over the entire area);
- *The area with no overlap* - the fraction of the total area with one serving AP and no other signals that are strong enough to interfere or contend with the server;
- *The area with one or more overlapping APs* - the fraction of the area where the serving AP overlaps with at least one AP (this figure sums up to one together with the fraction of the area with no overlap and coverage loss);
- *The area with two or more overlapping APs* - the fraction of the area where the serving AP overlaps with two or more APs.

Comparing coverage loss in the two tables, we observe that all the optimized solutions achieve in a significant improvement of this performance metric over that in the reference scenario. Thus, coverage loss reduces from 11.5 % to below 2 % in most of the solutions. In fact, coverage loss decreases when α increases within the first half of the unit interval. In the second half of the interval, coverage loss increases with α up to 3.99 % in Table 7.3 and up to 3.56 % in Table 7.4. The main reason for the improvement in all solutions compared to the reference scenario is that APs are not placed any more in service rooms in the inner part of the building where massive reinforced concrete walls obstruct the signal propagation. Unless overlap is emphasized, coverage improvement with growing α is explained by that APs are forced to be more evenly distributed over the two floors. With emphasized overlap minimization, however, they tend to group by three APs with only non-overlapping channels assigned within the group. This is the reason for the increasing coverage loss.

Throughput can be significantly improved compared to the reference scenario, as has been shown in Section 7.4.2, if AP locations are optimized with respect to the average net throughput. Although the throughput slightly decreases when the other optimization goal is also taken into account, the found solutions still give a significant improvement over the

Table 7.3: Performance of network designs for three non-overlapping channels

Performance metric		Reference scenario	Trade-off parameter α						
			0.0*	0.1 0.2	0.3 0.4 0.5	0.6 0.7	0.8	0.9	1.0
No overlap	[Area %]	47.57	60.74	66.82	77.73	78.44	78.26	77.64	72.38
Overlap $\geq$ 1 APs	[Area %]	40.93	37.55	31.56	21.41	20.59	18.30	18.37	24.06
Overlap $\geq$ 2 APs	[Area %]	4.00	3.50	4.52	5.18	4.55	3.74	3.99	3.43
Coverage loss	[Area %]	11.50	1.71	1.62	0.86	0.98	3.44	3.99	3.56
Av. net throughput	[Mbps]	10.69	13.16	13.16	13.10	13.06	12.66	12.41	12.19
W1-NT objective		10.84	13.30	13.28	13.15	13.03	12.55	12.33	12.16
W1-WCO objective	[$\times 10^6$]	12.79	6.20	2.11	1.39	1.23	0.95	0.87	0.87

* Sequential optimization

Table 7.4: Performance of network designs for three overlapping channels

Performance metric		Reference scenario	Trade-off parameter α					
			0.0*	0.1 0.2	0.3 0.4 0.5 0.6	0.7	0.8	0.9 1.0
No overlap	[Area %]	47.57	60.74	66.82	77.73	78.44	76.75	72.38
Overlap $\geq$ 1 APs	[Area %]	40.93	37.55	31.56	21.41	20.59	19.79	24.06
Overlap $\geq$ 2 APs	[Area %]	4.00	3.50	4.52	5.18	4.55	4.43	3.43
Coverage loss	[Area %]	11.50	1.71	1.62	0.86	0.98	3.45	3.56
Av. net throughput	[Mbps]	10.69	13.16	13.16	13.10	13.06	12.52	12.19
W1-NT objective		10.84	13.30	13.28	13.15	13.03	12.46	12.16
W1-WCO objective	[$\times 10^6$]	13.77	6.92	2.56	1.88	1.76	1.43	1.38

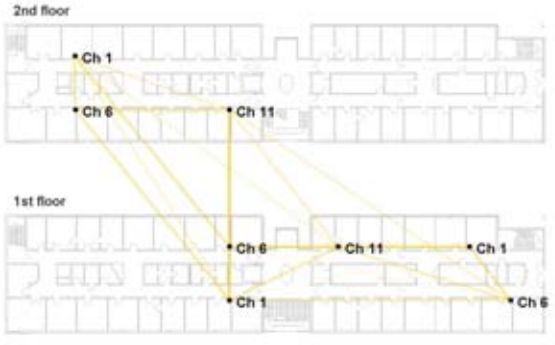
* Sequential optimization

reference scenario. Our numerical experiments show that integrating the AP placement model with the channel assignment decrease performance in terms of throughput, but the change is significant only for high α. Thus, as can be seen in both tables, up to $\alpha = 0.7$, throughput is hardly sacrificed when pursuing the other optimization goal, while a noticeable decrease can be observed only beyond this point.

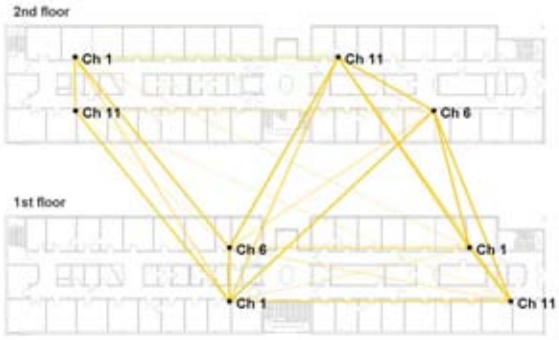
To achieve high network performance, the areas, where two or more APs operating on the same channel or having small channel distance overlap, are to be kept as small as possible. From the results, we observe that all the solutions obtained by the integrated model significantly outperform the reference scenario in terms of overlap. Furthermore, for small α, the size of the areas with no AP overlap is about twice as large than those with at least one overlapping AP. The maximum is achieved at $\alpha = \{0.6, 0.7\}$ and 0.7 for non-overlapping and overlapping channels, respectively. For larger α, the areas with no overlap and the areas with two or more overlapping APs slightly decrease in size, while the areas with only one overlapping AP slightly decrease.

Next, we analyze the optimized network designs with the focus on how the trade-off parameter affects AP distribution over the area and channel assignment in the optimal solution.

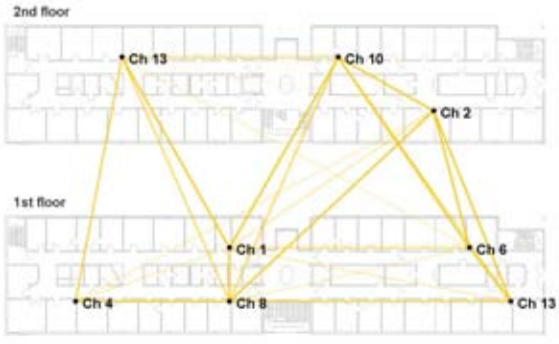
Minimum overlap solution ($\alpha = 1.0$). Figure 7.6(a) depicts the configuration with three non-overlapping channels and $\alpha = 1.0$ This is the configuration with emphasized overlap minimization. It is thus not surprising that APs are located in groups of three such that the set of all three non-overlapping channels is fully utilized within each such group.



(a) $\alpha = 1.0$, 3 channels: $\mathcal{C} = \{1, 6, 11\}$, $\mathcal{D} = \{0\}$



(b) $\alpha = 0.6$, 3 channels: $\mathcal{C} = \{1, 6, 11\}$, $\mathcal{D} = \{0, 5, 10\}$



(c) $\alpha = 0.6$, 13 channels: $\mathcal{C} = \{1, \ldots, 13\}$, $\mathcal{D} = \{0, \ldots, 12\}$

Figure 7.6: Overlap graphs for different solutions obtained by W1-NTWCO.

Throughput solution ($\alpha = 0.0$). An example of such network configuration has been obtained by the sequential approach discussed Section 7.4.2. In the configuration presented in Figure 7.7, five out of eight APs are placed on the first floor, which is expected because better coverage and net user throughput in some area on the adjacent floor is achieved if an AP is placed on the first floor. As of channel assignment, it is possible to distinguish groups of APs within which non-overlapping channels are used, although the grouping is less obvious than in the minimum overlap solution, and each group is

spread over both floors.

Trade-off solutions ($\alpha = 0.6$). A solution for three non-overlapping channels when $\alpha = 0.6$ is shown in Figure 7.9(c). The solution is characterized by the largest area with no overlap, high net user throughput, and at the same time relatively low coverage loss. This is seems to the best "fair" trade-off. We also observe an interesting pattern in the AP location and channel assignment. The APs are located in groups of two in a chess-board order among the floors. In one of these groups the channel distance is at maximum.

In the solution obtained for three overlapping channels W1-NTWCO (see Figure 7.6(b)), the AP placement just slightly differs from that for non-overlapping channels, although the channel assignment is clearly better from the interference point of view. This is because non-overlapping channels are not distinguished by the model, which does not cause any problem if APs are spatially well distributed, although non-overlapping channels may result in high interference in the area (both from MTs and APs) if APs are close to each other.

The optimal network design for 13 channels considering all of them as the overlapping is demonstrated in Figure 7.6(c). We observe an AP location pattern similar to that for the emphasized throughput, which is expected because, due to many available channels, the co-channel overlap is completely avoided, and the adjacent channel interference is at small (note that the minimum channel distance between APs that have considerable to some extent amount of overlap is four).

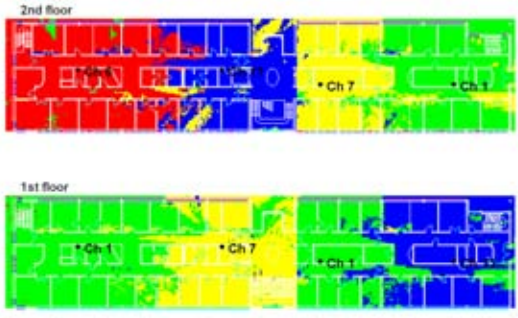
The numerical experiments reveal that the traditional sequential decision process of first deciding on AP positions and then assigning channels can be outperformed by joint optimization. The best mix of performance figures in our case is achieved when α is 0.6 or 0.7 for three non-overlapping channels and 0.7 for three overlapping channels. For these values, a significant reduction of overlap and contention can be achieved with only a marginal reduction of throughput. Interestingly, for smaller α values, but not below 0.3, the performance figures do not differ significantly from those for $\alpha = 0.6$ or $\alpha = 0.7$, which proves that the obtained solutions are robust. For higher values of α, solution quality degrades in all measures under examination.

7.5 Conclusions

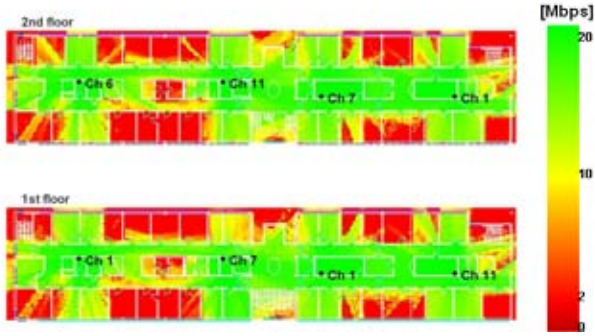
In this chapter, there have been presented models that can be utilized for effective planning and optimization of WLANs based on the IEEE 802.11 standard and operating in the infrastructure mode. The models address two design issues, optimization of AP locations and optimization of channel assignment. To optimize AP locations, user net throughput is used as a performance metric, whilst for channel assignment optimization there has been utilized the overlap measure reflecting the effects of potential contention as well as co- and adjacent channel interference. The standalone models for maximizing net user throughput and minimizing overlap have been presented first, and then combined into an integrated model.

In practice, it is usual to either apply optimization but handle both aspects sequentially or to adopt heuristic decisions. In this chapter, however, it has been demonstrated that a model that jointly optimizes the two decisions delivers solutions with better performance characteristics than those obtained by sequential optimization. Furthermore, the presented integrated model allows for controlling the relative priorities of the two objectives through the use of a trade-off parameter. A computational study conducted on realistic data has shown that when optimizing AP location and channel assignment jointly in a well-balanced way, a substantial reduction of channel overlap can be achieved at a price of a hardly noticeable reduction in throughput or coverage.

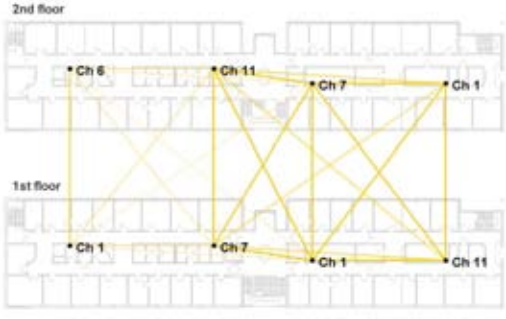
One of the interesting extensions of this work would be to verify the findings in a larger realistic planning scenario and to experiment with different traffic patterns by means of static and dynamic simulations. Another possible extension is to refine the model in order to also take into account the UL direction. A framework addressing this aspect is presented in the next chapter.



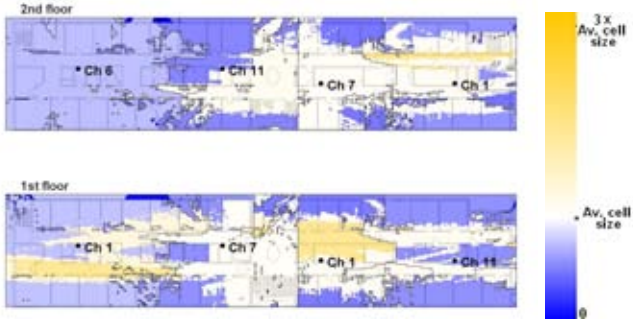
(a) Best server map



(b) User net throughput (average: 10.69 Mbps; areas with net throughput below 1 Mbps: 22.75 %, coverage loss: 11.5 %)

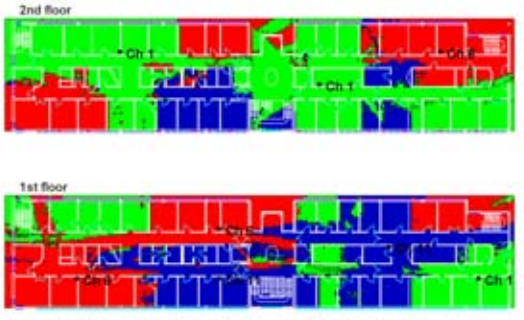


(c) Overlap graph

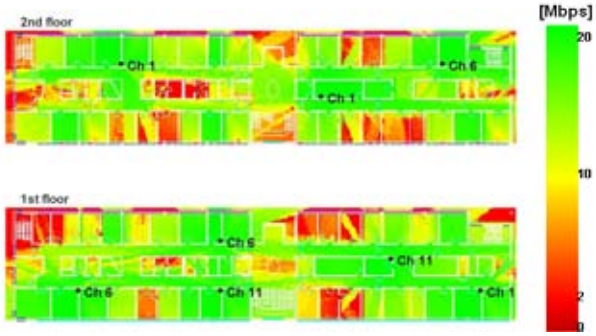


(d) Contention map

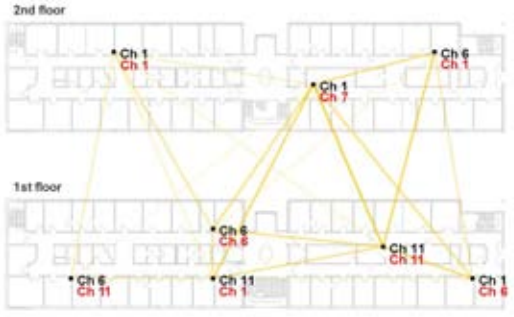
Figure 7.7: Reference scenario.



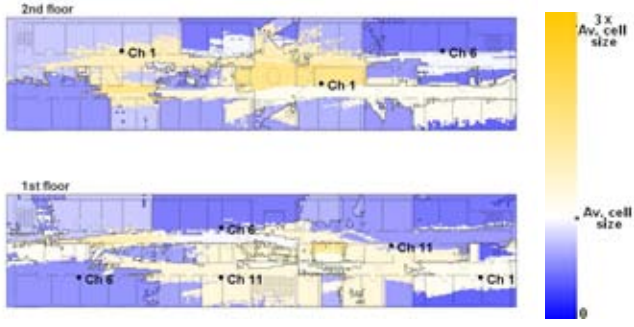
(a) Best server map



(b) User net throughput (average: 13.16 Mbps; areas with net throughput below 1 Mbps: 6.48 %, coverage loss: 1.71 %)

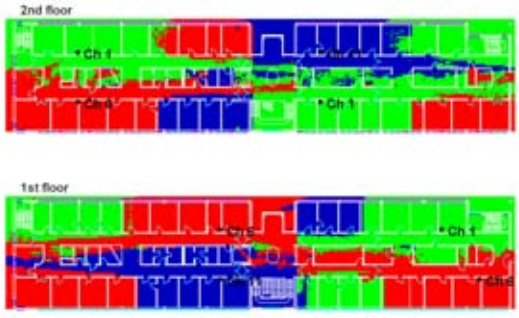


(c) Overlap graph (four-channel assignment is shown in red)

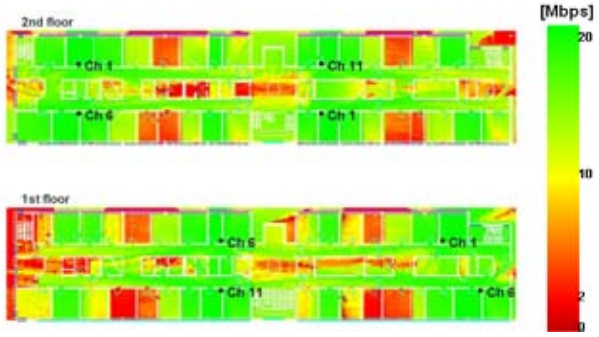


(d) Contention map

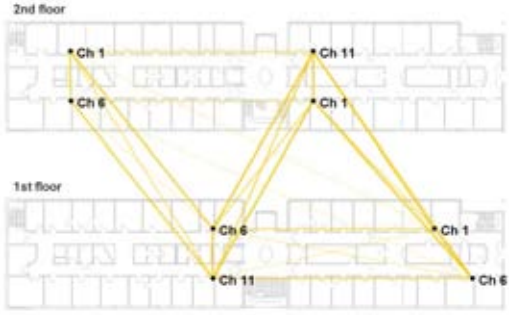
Figure 7.8: Sequential optimization (W1-NT, W1-WCO with 3 non-overlapping channels).



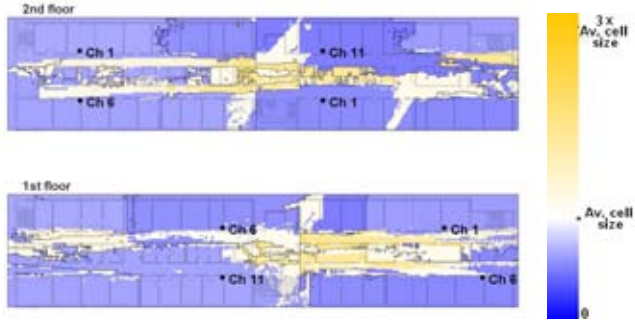
(a) Best server map



(b) User net throughput (average: 13.05 Mbps, areas with net throughput below 1 Mbps: 5.23 %, coverage loss: 0.98 %)



(c) Overlap graph



(d) Contention map

Figure 7.9: Joint optimization (W1-NTWCO with 3 non-overlapping channels), $\alpha = 0.6$.

Chapter 8

A Contention-aware Optimization Model for Channel Assignment and AP Transmit Power Adjustment

In this chapter, we present a framework for joint optimization of channel assignment and AP transmit power taking into account potential contention among user terminals. Similar to models presented Chapter 7, the approach and the model presented in the current chapter are applicable to any type of IEEE 802.11 network operating in the infrastructure mode, although the contention issue is probably less important for IEEE 802.11a. In numerical experiments, we focus on IEEE 802.11g networks which are more common for office environments.

8.1 System Model

The system model used in this chapter extends the one presented in Section 7.1. The main difference between the two system models is that here we take also into account UL. In addition to this, now a CS range may also include APs. Therefore, we redefine the CS range as follows,

$$\mathcal{R}_{il}^{cs} = \{j \in \mathcal{J} \cup \mathcal{A} \setminus \{i\} : l \cdot g_{ij} \geq \gamma^{cs}\}, \tag{8.1}$$

where l is the transmit power level of STA i, and g_{ij} is the power gain between STA i and STA j (recall that a STA can be either an MT or an AP).

When uniform AP transmit power is assumed (i.e., the coverage range of each AP is approximately the same), minimizing the AP overlap is a straightforward strategy in frequency assignment problems for WLANs. This strategy was considered in Chapter 7. However, when the transmit power varies by AP, channel assignment optimization driven by minimizing overlap areas between APs may result in network performance degradation caused by increased interference and contention. This is because the coverage range of an MT can be larger than that of an AP, which often happens since the transmit power level of an MT is usually user-controlled and is therefore typically set at its maximum level.

A situation where a channel assignment based on minimizing AP overlap increases interference or, in the worst case, contention probability is demonstrated in Figure 8.1. In the figure, a solid line denotes the serving range of an AP, dashed line corresponds to the CS range of an MT (j or j'), and a dash-dot line marks an AP's range considered for overlap estimation, i.e., the AP coverage overlap is defined as the size of the area where the ranges of two APs intersect. (Both the CS range and the interference range can be used for overlap calculation, although considering the CS range is more common when avoiding contention is prioritized over interference reduction.) Note that circular ranges in the figure are used merely to simplify the presentation.

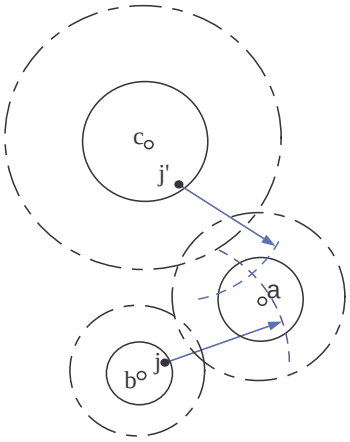


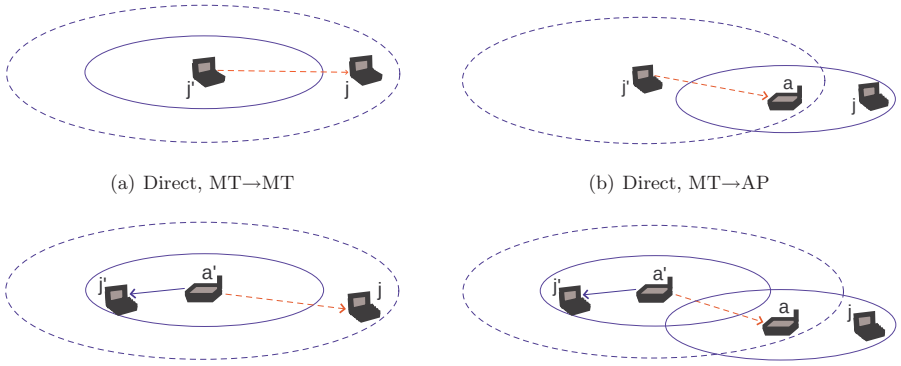
Figure 8.1: Channel assignment driven by minimizing overlap areas between APs: AP b having a smaller overlap with AP a serves MTs that may contend with AP a.

Consider a set of adjacent (interfering) channels for this example. If the overlap weighted with some channel-distance dependent coefficient $F(d)$ is minimized (see, for example, Section 7.2.2), the channel distance between a pair of APs (a, b) is likely to be smaller than that for (a, c) because the overlap area of the former is smaller. On the other hand, since the MT transmit power does not depend on the AP transmit power and is the same for MTs assigned to AP b and those assigned to AP c, the interference coming from MTs assigned to b is more likely to be higher than that from MTs assigned to c (for the same channel distance for both pairs (a, b) and (a, c)). Thus, a smaller channel distance for (a, b) further increases the interference in the serving area of AP a coming from the serving area of AP b. Moreover, if a channel distance of zero is chosen for (a, b) and a non-zero channel distance is chosen for (a, c), then the channel assignment may cause contention between MTs in the serving areas of AP b and AP a. Therefore, from the contention point of view, if there is no other channel available, having zero channel distance for (a, c) and a non-zero channel distance for (a, b) would be a better solution in this situation. The example demonstrates that when channel assignment is combined with AP transmit power control, transmit power of MTs needs to be also taken into account. Addressing this issue, we use a contention metric.

Similar to [51], we distinguish four possible contention types depicted in Figure 8.2. (Recall that contention occurs when STAs operate on the same channel.) We say that MT j' is a direct contender of MT j if either MT j or its serving AP a is within the CS range of MT j', provided that both MTs operate on the same channel. These two scenarios are demonstrated in Figures 8.2(a) and 8.2(b), respectively. MT j' contends with MT j indirectly, when j' communicates with its serving AP a' and this blocks communication between MT j and its serving AP a (see Figures 8.2(c) and 8.2(d)). In a simplified way, the symbolic notation of the four situations can be read as follows,

- MT→MT: MT j' contends with MT j,
- MT→AP: MT j' contends with the AP serving MT j,
- AP(MT)→MT: when serving MT j', the AP contends with MT j,
- AP(MT)→AP: when serving MT j', the AP contends with the AP serving MT j.

In our model, we consider direct and indirect contention to serving APs (see Figures 8.2(b) and 8.2(d), respectively). We define a set of parameters $\nu^l_{ab}, \{a, b\} \subseteq \mathcal{A}, l \in \mathcal{L}$, such



(a) Direct, MT→MT

(b) Direct, MT→AP

(c) Indirect, AP(MT)→MT

(d) Indirect, AP(MT)→AP

Figure 8.2: Contention types.

that ν^l_{ab} represents the size of the area (the number of TPs) where each TP is within the serving range of AP a and either of the following conditions holds,

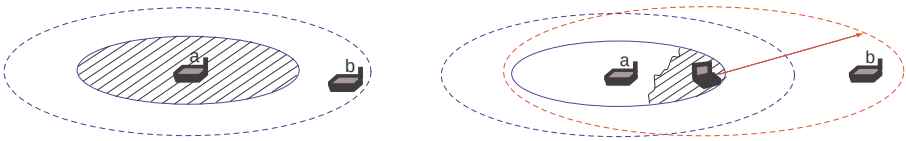
- AP a covers with its CS range AP b (indirect contention, AP(MT)→AP),
- AP b is within the CS range of an MT located at the TP (direct contention, MT→AP).

The two cases are shown in Figures 8.3(b) and 8.3(a). In each of the two figures, the size of the shaded area equals the corresponding parameter ν^l_{ab}. Mathematically, the parameters are defined as follows,

$$\nu^l_{ab} = \begin{cases} |\mathcal{R}^{srv}_{al}| & \text{if } b \in \mathcal{R}^{cs}_{al}, \\ |\{j \in \mathcal{R}^{srv}_{al} : b \in \mathcal{R}^{cs}_{j\,L_{MT}}\}| & \text{otherwise,} \end{cases} \tag{8.2}$$

where L_{MT} is the transmit power level of MT j. Note that ν^l_{ab} may be equal to $|\mathcal{R}^{srv}_{al}|$ even when $b \notin \mathcal{R}^{cs}_{al}$. This happens when AP b is within the CS range of any MT within the serving range of AP a. The situation is very likely to occur when MT's transmit power is not controlled and therefore set to the maximum level which results in that MTs have larger CS ranges than APs.

We aim at minimizing potential contention areas by adjusting AP transmit power levels and finding a proper channel assignment such that full coverage of a given area is maintained. Moreover, we are interested in a model that would allow us to also deal with overlapping channels.



(a) AP b is within the CS range of AP a: contention parameter represents the entire serving area of AP a.

(b) AP b is outside the CS range of AP a: contention parameter represents a part of the serving range of AP a from where an MT can cover AP b with its CS range.

Figure 8.3: Modeling contention parameter ν^l_{ab}.

8.2 Optimization Problem

To formulate an optimization problem, we introduce the following three sets of (binary) decision variables, power variables $\{p_a^l, a \in \mathcal{A}, l \in \mathcal{L}\}$ modeled as

$$p_a^l = \begin{cases} 1 & \text{if AP } a \text{ transmits at power level } l, \\ 0 & \text{otherwise,} \end{cases}$$

channel variables $\{f_a^c, a \in \mathcal{A}, c \in \mathcal{C}\}$ modeled as

$$f_a^c = \begin{cases} 1 & \text{if AP } a \text{ uses channel } c, \\ 0 & \text{otherwise,} \end{cases}$$

contention variables $\{y_{ab}^{ld}, \{a, b\} \subseteq \mathcal{A}, l \in \mathcal{L}, d \in \mathcal{D}\}$ defined as

$$y_{ab}^{ld} = \begin{cases} 1 & \text{if AP } a \text{ uses power level } l, \text{ and } d \text{ is the channel distance for APs } a \text{ and } b, \\ 0 & \text{otherwise.} \end{cases}$$

To strengthen the model (the LP relaxation of its frequency assignment part, in particular), we also use a set of coupling variables $\{s_{ab}^{c_1c_2}, (a, b) \in \mathcal{A}^2, c_1, c_2 \in \mathcal{C}\}$, where $\mathcal{A}^2 = \{(a, b) : a, b \in \mathcal{A}, a < b\}$, defined as follows,

$$s_{ab}^{c_1c_2} = \begin{cases} 1 & \text{if AP } a \text{ uses channel } c_1 \text{ and AP } b \text{ uses channel } c_2, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to minimize the sum of weighted contention/interference estimation for all pairs of APs. Based on our preliminary contention study, we choose to focus on direct and indirect contention with serving APs, i.e., contention types demonstrated in Figures 8.2(b) and 8.2(d). (It will be demonstrated later by examples in Section 8.5 that the decision is reasonable.)

To allow for overlapping channels, we use weighting coefficients in the objective function (similar to models W1-WCO and W1-NTWCO). In our numerical experiments function $F(d)$ is defined as described in Section 7.2.2. Note that d may equal zero, if two APs operate on the same channel. The weighted sum of contention metrics allows us to estimate the found network configuration. The use of the contention metric for AP pairs with channel distance $d > 0$ is justified by the fact that it correlates with interference.

The objective function is designed in a way that a higher priority is given to solutions in which neighboring APs with potentially large contention areas have larger channel distance. This allows us to reduce not only contention probability but also possible adjacent channel interference. The mathematical formulation (denoted by W2-CP) is presented below.

$$\text{[W2-CP]} \quad \sum_{\{a,b\}\subseteq\mathcal{A}} \sum_{l\in\mathcal{L}} \sum_{d\in\mathcal{D}} \frac{\nu_{ab}^{l}}{(1+d)^{k}} \cdot y_{ab}^{ld} \longrightarrow \min \tag{8.3a}$$

$$\text{s. t.} \quad \sum_{l\in\mathcal{L}} p_a^l = 1 \qquad a \in \mathcal{A} \tag{8.3b}$$

$$\sum_{l\in\mathcal{L}} \sum_{\substack{a\in\mathcal{A}:\\ j\in\mathcal{R}_{al}^{srv}}} p_a^l \geq 1 \qquad j \in \mathcal{J} \tag{8.3c}$$

$$\sum_{c\in\mathcal{C}} f_a^c = 1 \qquad a \in \mathcal{A} \tag{8.3d}$$

$$f_a^{c_1} = \sum_{c_2\in\mathcal{C}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, c_1 \in \mathcal{C} \tag{8.3e}$$

$$f_b^{c_2} = \sum_{c_1\in\mathcal{C}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, c_2 \in \mathcal{C} \tag{8.3f}$$

$$\sum_{l\in\mathcal{L}} y_{ab}^{ld} = \sum_{\substack{c_1,c_2\in\mathcal{C}:\\ |c_1-c_2|=d}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, d \in \mathcal{D} \tag{8.3g}$$

$$\sum_{l\in\mathcal{L}} y_{ba}^{ld} = \sum_{\substack{c_1,c_2\in\mathcal{C}:\\ |c_1-c_2|=d}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, d \in \mathcal{D} \tag{8.3h}$$

$$\sum_{d\in\mathcal{D}} y_{ab}^{ld} = p_a^l \qquad \{a,b\} \subseteq \mathcal{A}, l \in \mathcal{L} \tag{8.3i}$$

$$p_a^l \in \{0,1\} \qquad a \in \mathcal{A}, l \in \mathcal{L} \tag{8.3j}$$

$$f_a^c \in \{0,1\} \qquad a \in \mathcal{A}, c \in \mathcal{C} \tag{8.3k}$$

$$y_{ab}^{ld} \in \{0,1\} \qquad \{a,b\} \subseteq \mathcal{A}, l \in \mathcal{L}, d \in \mathcal{D} \tag{8.3l}$$

$$s_{ab}^{c_1c_2} \in \{0,1\} \qquad (a,b) \in \mathcal{A}^2, c_1, c_2 \in \mathcal{C} \tag{8.3m}$$

By constraints (8.3j)-(8.3m), all variables in the model are binary. Constraints (8.3b) ensure that exactly one transmit power level is selected for each AP. By constraints (8.3c), for any TP there is at least one AP that uses a power level such that the TP is within the serving range of the AP. In other words, every TP must have at least one potential server. Note that constraints (8.3c) model a full coverage requirement. (Partial coverage can be modeled by modifying the right-hand sides in constraints (8.3c) and introducing an additional overall coverage constraint by which a certain coverage degree in the network is achieved.)

By constraints (8.3d), exactly one channel must be selected for each AP. If a channel is not used by an AP, all corresponding coupling variables must be zero. This is modeled by equalities (8.3e) and (8.3f). Constraints (8.3g) and (8.3h) set relation between contention variables and coupling variables, and constraints (8.3i) ensure that if an AP is within the CS range of another AP for some chosen transmit power level, then the two APs either contend ($d = 0$) or interfere ($d > 0$). Here, we assume that any channel interferes to any other, i.e., $\mathcal{D}$ includes all possible channel distances for the two APs, although the model can be easily adapted to a case when transmissions on two channels interfere only if the channel distance does not exceed some $\bar{d}$. That is, instead of $\mathcal{D}$ we use in the model subset $\mathcal{D}' = \{d \in \mathcal{D} : d \leq \bar{d}\}$ and change constraints (8.3i) to $\sum_{d\in\mathcal{D}'} y_{ab}^{ld} \leq p_a^l, \{a,b\} \subseteq \mathcal{A}, l \in \mathcal{L}$.

Observe that the model can be easily extended to also enable decisions on AP locations, i.e., choosing APs to install from a given set of candidate locations. One of the possible modifications to handle this is changing constraints (8.3b) and (8.3d) to inequalities with (binary) installation variables in the right-hand side. In this chapter, the set of installed APs

is assumed to be known. This is mainly for the reason that network reconfiguration is usually performed more often than changing network topology.

The theoretical computational complexity of problem W2-CP is formalized in Theorem 8.1.

Theorem 8.1. *W2-CP is $\mathcal{NP}$-hard.*

Proof. An optimization problem is $\mathcal{NP}$-hard if it has an $\mathcal{NP}$-complete recognition version, i.e., the corresponding recognition problem (let it be denoted by W2-CP-r) is in the class $\mathcal{NP}$, and all other problems in $\mathcal{NP}$ polynomially transform to W2-CP-r.

[W2-CP-r] *Does there exist a network configuration with the total contention penalty of at most C^*?*

W2-CP-r is in $\mathcal{NP}$ since every yes instance of W2-CP-r can be verified in polynomial time. To prove that all other problems in $\mathcal{NP}$ polynomially transform to W2-CP-r, it is sufficient to show that a known $\mathcal{NP}$-complete problem polynomially transforms to W2-CP-r. We consider a minimum-cost set covering problem (MCSCP). The problem is $\mathcal{NP}$-hard since it has as its special case the minimum set covering problem which is known to be $\mathcal{NP}$-hard. We show that a recognition version of a minimum-cost set covering problem (let it be denoted by MCSCP-r) polynomially transforms to W2-CP-r, i.e., for every instance $\mathrm{I}_{\text{MCSCP-r}}$ of the former we can construct in polynomial time an instance $\mathrm{I}_{\text{W2-CP-r}}$ of W2-CP-r such that $\mathrm{I}_{\text{MCSCP-r}}$ is a yes instance of the MCSCP-r if an only if $\mathrm{I}_{\text{W2-CP-r}}$ is a yes instance of W2-CP-r.

Consider a recognition version of an MCSCP with the set of elements $\mathcal{J}$, set of subsets $\{\mathcal{S}_i \subseteq \mathcal{J}, i \in \mathcal{I}\}$, set of costs $\{c_i, i \in \mathcal{I}\}$, where c_i is a cost of choosing subset $\mathcal{S}_i$, and maximum cost C^*. The mathematical formulation of the MCSCP-r is as follows.

$$\text{[MCSCP-r]} \quad \sum_{i \in \mathcal{I}} c_i x_i \leq C^* \tag{8.4a}$$

$$\sum_{\substack{i \in \mathcal{I}: \\ j \in \mathcal{S}_i}} x_i \geq 1 \qquad j \in \mathcal{J} \tag{8.4b}$$

$$x_i \in \{0,1\} \qquad i \in \mathcal{I} \tag{8.4c}$$

The corresponding instance of W2-CP-r has a set of APs $\mathcal{A}$ (the set coincides with set $\mathcal{I}$ in the MCSCP-r), a set of power levels available at each AP $\mathcal{L} = \{l_0, l_1\}$, and a set of TPs $\mathcal{J}'$ (that contains all the elements from set $\mathcal{J}$ in the MCSCP and $|\mathcal{A}|$ unique elements that do not belong to $\mathcal{J}$, i.e., $\mathcal{J}' = \mathcal{J} \cup \{j'_1, \ldots, j'_{|\mathcal{A}|}\}$). Consider a network with a single channel available. This implies that set $\mathcal{D}$ consists of one element (zero). With this assumption, the channel assignment part of model W2-CP can be discarded since all APs operate on the same channel. Thus, constraints (8.3d)–(8.3h) are not needed and can be simply omitted. The y-variables can be replaced by the p-variables and constraints (8.3i) can be brought into the objective function as follows,

$$\sum_{(a,b) \in \mathcal{A}^2} \sum_{l \in \mathcal{L}} \sum_{d \in \mathcal{D}} \frac{\nu_{ab}^l}{(1+d)^k} \cdot y_{ab}^{ld} = \sum_{(a,b) \in \mathcal{A}^2} \sum_{l \in \mathcal{L}} \nu_{ab}^l p_a^l = \sum_{a \in \mathcal{A}} \sum_{\substack{b \in \mathcal{A}: \\ b \neq a}} \sum_{l \in \mathcal{L}} \nu_{ab}^l p_a^l =$$

$$\sum_{a \in \mathcal{A}} \sum_{l \in \mathcal{L}} \left[\left(\sum_{\substack{b \in \mathcal{A}: \\ b \neq a}} \nu_{ab}^l \right) p_a^l \right].$$

Now we show that any instance of the MCSCP-r can be polynomially transformed into the described instance of W2-CP-r. Given an instance of the MCSCP-r, we create a set of

ranges $\{\mathcal{R}^{srv}_{a,l}, a \in \mathcal{A}, l \in \mathcal{L}\}$, and a set of contention parameters such that for any AP a the following conditions hold,

$$\begin{aligned}
&\mathcal{R}^{srv}_{a,l_0} = \{j'_a\},\\
&\mathcal{R}^{srv}_{a,l_1} = \mathcal{S}_a \cup \{j'_a\},\\
&\nu^{l_0}_{ab} = \varepsilon, \qquad \forall b \in \mathcal{A} \setminus \{a\},\\
&\nu^{l_1}_{ab} \geq 0, \qquad \forall b \in \mathcal{A} \setminus \{a\}\\
&\sum_{\substack{b\in\mathcal{A}:\\ b\neq a}} \nu^{l_1}_{ab} = c_a,
\end{aligned}$$

where $\mathcal{S}_a = \mathcal{S}_i,\ \forall a = i, a \in \mathcal{A}, i \in \mathcal{I}$ (recall that $\mathcal{A} = \mathcal{I}$ by definition), and ε is a small positive number.

The formulation for the resulting instance of W2-CP-r is as shown below,

$$\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{L}}\left[\left(\sum_{\substack{b\in\mathcal{A}:\\ b\neq a}} \nu^{l}_{ab}\right) p^l_a\right] \leq C^* + |\mathcal{A}| \cdot (|\mathcal{A}| - 1) \cdot \varepsilon \tag{8.5a}$$

$$\sum_{l\in L}\sum_{\substack{a\in\mathcal{A}:\\ j\in\mathcal{R}^{srv}_{a,l}}} p^l_a \geq 1, \quad j \in \mathcal{J} \tag{8.5b}$$

$$p^l_a \in \{0,1\}, \qquad a \in \mathcal{A}, l \in \mathcal{L} \tag{8.5c}$$

Note that constraints (8.3b) are omitted since they are satisfied by the instance construction.

The transformation can be clearly done in polynomial time. By construction, it is true that any instance of the MCSCP-r is a yes instance if and only if the constructed instance of W2-CP-r is. Thus, we have showed that the MCSCP-r polynomially transforms to W2-CP-r. Hence the conclusion. □

8.3 Lower Bounding

In this section, we investigate several approaches for finding a lower bound to problem W2-CP. One straightforward way to obtain a lower bound to W2-CP is to solve its LP relaxation. The LP-relaxation solution to W2-CP can be obtained very fast. However, it is very weak, i.e., gives a very poor lower bound on the optimal objective function value. To exploit the computational efficiency of solving LP-relaxation, we strengthen the LP version of W2-CP by introducing additional constraints that apply to the channel assignment part of the problem. Two lower bounding approaches based on strengthening the LP relaxation are presented in Section 8.3.1 and Section 8.3.2. In Section 8.3.3, we present the third approach which exploits the idea of decomposing the (integer) problem into the power assignment part and the channel assignment part.

8.3.1 Approach 1

The approach is based on the following observation. Due to a small number of available channels in IEEE 802.11 spectrum and even fewer non-overlapping channels, the number of APs using the same channel is large (compared to the total number of APs) which results in many pairs where both APs use the same channel. Furthermore, not so many parameters ν^l_{ab} in the objective function of W2-CP are non-zero because of the large gap between the serving threshold and the CS threshold. This suggests that introducing a constraint setting the minimum number of pairs of contending APs would improve the LP bound of the problem. Let θ denote the lower bound on the minimum number of AP pairs in the network operating on

the same channel. With the notation used in model W2-CP, the constraint can be formulated as follows,

$$\sum_{(a,b)\in\mathcal{A}^2}\sum_{c\in\mathcal{C}} s_{ab}^{cc} \geq \theta \ . \tag{8.6}$$

The first approach is thus to solve the LP relaxation to W2-CP with constraint (8.6). We further prove that the value of θ can be derived theoretically. This is formulated in Theorem 8.2. The result of the theorem relates to the well-known Turán's theorem [10, 64], one of the fundamental theorems in Extremal Graph Theory, and was also recognized, for example, in the context of frequency assignment problems for GSM [21]. For the sake of completeness we provide a formal proof below.

Theorem 8.2. *The minimum number of AP pairs where both APs are assigned the same channel is*

$$\theta = 0.5 \cdot \left\lfloor \frac{|\mathcal{A}|}{|\mathcal{C}|} \right\rfloor \cdot (|\mathcal{A}| - |\mathcal{C}| + r),$$

where r is the remainder of dividing the number of APs by the number of available channels, i.e., $r = |\mathcal{A}|\%|\mathcal{C}|$.

Proof. Let us introduce additional notation. Let $\mathcal{S}_c$ denote a set of APs using channel c, n_c be the size of the set and $\mathbf{n}$ be a vector of elements $n_c, c \in \mathcal{C}$. Then, the number of AP pairs operating on channel c is given by the binomial coefficient $\binom{n_c}{2} = \frac{n_c\cdot(n_c-1)}{2}$. Thus, in the entire network, the number of AP pairs using the same channel is $F = \sum_{c\in\mathcal{C}} \binom{n_c}{2}$. To find the minimum possible F, we formulate the following optimization problem (we denote this problem P1).

[P1] *Find a partition of $\mathcal{A}$ such that $\bigcup_{c\in\mathcal{C}} \mathcal{S}_c = \mathcal{A}$ and $F = \sum_{c\in\mathcal{C}} \frac{n_c\cdot(n_c-1)}{2}$ is minimized.*

Consider a simple transformation,

$$F = \sum_{c\in\mathcal{C}} \frac{n_c \cdot (n_c - 1)}{2} = 0.5 \cdot \sum_{c\in\mathcal{C}} (n_c^2 - n_c) =$$
$$0.5 \cdot (\sum_{c\in\mathcal{C}} n_c^2 - \sum_{c\in\mathcal{C}} n_c) = 0.5 \cdot (\sum_{c\in\mathcal{C}} n_c^2 - |\mathcal{A}|) = 0.5 \cdot (f(\mathbf{n}) - |\mathcal{A}|) \ , \tag{8.7}$$

where $f(\mathbf{n}) = \sum_{c\in\mathcal{C}} n_c^2$. Observe that F achieves its minimum when $f(\mathbf{n})$ is minimum.

Let us minimize $f(\mathbf{n})$ subject to $\sum_{c\in\mathcal{C}} n_c = |\mathcal{A}|$ and $n_c \geq 0, c \in \mathcal{C}$, and let the optimization problem be denoted by P2. Observe that if n_c-variables are not constrained to be integral, the optimal solution to P2 is $n_c = \frac{|\mathcal{A}|}{|\mathcal{C}|}, c \in \mathcal{C}$. We, however, are interested in an integer optimal solution which we derive below.

Let $\mathcal{C}'$ be a subset of $\mathcal{C}$ such that

$$n_c = \left\lfloor \frac{|\mathcal{A}|}{|\mathcal{C}|} \right\rfloor + 1 + \mu_c \ , \qquad \forall c \in \mathcal{C}' \tag{8.8a}$$
$$n_c = \left\lfloor \frac{|\mathcal{A}|}{|\mathcal{C}|} \right\rfloor - \nu_c \ , \qquad \forall c \in \mathcal{C} \setminus \mathcal{C}' \tag{8.8b}$$

where $0 \leq \mu_c \leq |\mathcal{A}| - 1 - \left\lfloor \frac{|\mathcal{A}|}{|\mathcal{C}|} \right\rfloor, c \in \mathcal{C}_1$ and $0 \leq \nu_c \leq \left\lfloor \frac{|\mathcal{A}|}{|\mathcal{C}|} \right\rfloor, c \in \mathcal{C} \setminus \mathcal{C}'$ are unknown integers. Note also that the size of set $\mathcal{C}'$ is not given. Since the number of all APs on all channels is

$|\mathcal{A}|$, the following equations are valid and equivalent,

$$\sum_{c\in\mathcal{C}'}\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor+1+\mu_c\right)+\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor-\nu_c\right)=|\mathcal{A}|\ , \tag{8.9}$$

$$|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor+|\mathcal{C}'|+\sum_{c\in\mathcal{C}'}\mu_c-\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\nu_c=|\mathcal{A}|\ , \tag{8.10}$$

$$|\mathcal{C}'|+\sum_{c\in\mathcal{C}'}\mu_c-\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\nu_c=r\ , \tag{8.11}$$

where $r=|\mathcal{A}|\%|\mathcal{C}|$. Furthermore,

$$f(\mathbf{n})=\sum_{c\in\mathcal{C}}n_c^2=\sum_{c\in\mathcal{C}'}\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor+(1+\mu_c)\right)^2+\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor-\nu_c\right)^2=$$

$$|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor^2+2\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot\left(\sum_{c\in\mathcal{C}'}(1+\mu_c)-\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\nu_c\right)+\sum_{c\in\mathcal{C}'}(1+\mu_c)^2+\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\nu_c^2=$$

$$|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor^2+2\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot r+\sum_{c\in\mathcal{C}'}(1+\mu_c)^2+\sum_{c\in\mathcal{C}\setminus\mathcal{C}'}\nu_c^2\ . \tag{8.12}$$

Observe that it follows from (8.12) that $f(\mathbf{n})$ achieves its minimum when $\mu_c=0,\forall c\in\mathcal{C}'$ and $\nu_c=0,\forall c\in\mathcal{C}\setminus\mathcal{C}'$. Note that this integer solution is feasible to P2 (it satisfies equation (8.11) when $|\mathcal{C}'|=r$) and thus also optimal. The optimal value of F is found below,

$$F=0.5\cdot(f(\mathbf{n})-|\mathcal{A}|)=0.5\cdot\left(\left(|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor^2+2\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot r+r\right)-|\mathcal{A}|\right)=$$

$$0.5\cdot\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot\left(|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor+2\cdot r\right)+r-|\mathcal{A}|)\right)=0.5\cdot\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot(|\mathcal{A}|+r)+r-|\mathcal{A}|\right)=$$

$$0.5\cdot\left(\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot(|\mathcal{A}|+r)-|\mathcal{C}|\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\right)=0.5\cdot\left\lfloor\frac{|\mathcal{A}|}{|\mathcal{C}|}\right\rfloor\cdot(|\mathcal{A}|-|\mathcal{C}|+r)\ . \tag{8.13}$$

Hence the conclusion. □

8.3.2 Approach 2

In the second approach, we subdivide the initial channel set into two sub-domains $\mathcal{C}'$ and $\mathcal{C}\setminus\mathcal{C}'$ and introduce an additional constraint,

$$\sum_{c\in\mathcal{C}'}f_a^c=\tau_a,\qquad a\in\mathcal{A}, \tag{8.14}$$

where $\tau_a, a\in\mathcal{A}$, is a binary variable which is one if and only if AP a is assigned one of the channels from sub-domain $\mathcal{C}'$, otherwise is zero. In other words, constraints (8.14) force all positive f-variables to belong to a subset of $\mathcal{C}$. The constraints are valid in W2-CP, since only one f-variable for any AP is positive (equals one) in an optimal integer solution to W2-CP.

By this approach, we solve a problem constructed from the LP-relaxation to W2-CP and constraints (8.14). Observe that only the τ-variables are integral in the resulting formulation.

In the optimal solution to the LP relaxation of W2-CP, the channel assignment at each AP tends to have fractional values for the most distant channels (i.e., channels 1 and 13, if both channels belong to the channel set $\mathcal{C}$) and zeros for the others. Channel grouping prohibits such decisions setting more restrictions on the optimal solution and thus improves the lower bound.

The approach can be extended for deriving a heuristic that further improves the obtained lower bound. A similar idea based on channel grouping was also exploited in [47] where the authors presented an iterative lower bounding procedure that produces a non-decreasing sequence of lower bounds. In each iteration, the authors solve a smaller subproblem in which each subset of frequencies is treated as a single frequency. The minimum of individual penalties of the frequencies in the subset is used as the penalty in the subproblem.

8.3.3 Approach 3

The third lower bounding approach considers integer model W2-CP (not its LP relaxation) and is based on an observation that model W2-CP actually consists of the power allocation part and the channel assignment part. In fact, a solution with any channel assignment and a power setting satisfying the full coverage constraint is feasible in W2-CP. Observe that a lower bound for W2-CP can be obtained by relaxing constraints (8.3i) and then solving the resulting problem where constraints (8.3b), (8.3c), and (8.3j) can also be omitted since they involve variables that appear neither in other constraints nor in the objective function.

The remaining problem (let it be denoted by W2-CP-CA) consists of the objective function (8.3a) and constraints (8.3d)-(8.3h) and (8.3k)-(8.3m). The problem can be treated as a pure channel assignment problem. This is formalized in Theorem 8.3.

Theorem 8.3. *W2-CP-CA can be reduced to channel assignment problem W1-WCO presented in Section 7.2.2.*

Proof. Observe that in the considered formulation the only remaining restrictions connected to power allocation are the ones defined by constraints (8.3g) and (8.3h). The constraints state that if the channel distance between two APs (a, b) equals d, there must be exactly one power level l defining a non-zero y-variable for the given combination of (a, b) and d. In the optimal solution, it is the power level of AP a having the smallest ν^l_{ab} for the given pair of APs (this can be seen from the objective function). Let w^d_{ab} be a binary variable which is one if APs a and b operate with channel distance d, and zero otherwise. Observe that it follows from constraints (8.3g) and (8.3h) and the definition of the w-variables and the y-variables that $w^d_{ab} = \sum_{l\in\mathcal{L}} y^{ld}_{ab} = \sum_{l\in\mathcal{L}} y^{ld}_{ba}$. Therefore, the objective of the considered formulation can be transformed as follows,

$$\sum_{\{a,b\}\subseteq\mathcal{A}} \sum_{l\in\mathcal{L}} \sum_{d\in\mathcal{D}} \frac{\nu^l_{ab}}{(1+d)^k} \cdot y^{ld}_{ab} = \sum_{(a,b)\in\mathcal{A}^2} \sum_{l\in\mathcal{L}} \sum_{d\in\mathcal{D}} \left(\frac{\nu^l_{ab}}{(1+d)^k} \cdot y^{ld}_{ab} + \frac{\nu^l_{ba}}{(1+d)^k} \cdot y^{ld}_{ba} \right) ,$$

which is equivalent to

$$\sum_{(a,b)\in\mathcal{A}^2} \sum_{d\in\mathcal{D}} \frac{\min_{l\in\mathcal{L}} \nu^l_{ab} + \min_{l\in\mathcal{L}} \nu^l_{ba}}{(1+d)^k} \cdot w^d_{ab} = \sum_{(a,b)\in\mathcal{A}^2} \sum_{d\in\mathcal{D}} \frac{v_{ab}}{(1+d)^k} \cdot w^d_{ab} ,$$

where $v_{ab} = \min_{l\in\mathcal{L}} \nu^l_{ab} + \min_{l\in\mathcal{L}} \nu^l_{ba}$ is a parameter.

With the new set of variables introduced above $\left(\{w^d_{ab}, (a, b) \in \mathcal{A}^2\}\right)$, constraints (8.3g) and (8.3h) can be substituted by the following set of constraints,

$$w^d_{ab} = \sum_{\substack{c_1,c_2\in\mathcal{C}:\\ |c_1-c_2|=d}} s^{c_1c_2}_{ab} \qquad (a, b) \in \mathcal{A}^2, d \in \mathcal{D} .$$

Constraints (8.3l) can now be replaced with integrality constraints on w-variables.

The new formulation does not involve power assignment and its optimal solution is also optimal for the model obtained from W2-CP by relaxing constraints (8.3i) and removing power constraints. Hence the conclusion. □

Corollary 8.1. *An optimal solution to W2-CP-CA gives a lower bound on the optimal value of the objective function of W2-CP.*

The complete formulation of W2-CP-CA is presented below.

$$\text{[W2-CP-CA]} \quad \sum_{(a,b)\in\mathcal{A}^2} \sum_{d\in\mathcal{D}} \frac{v_{ab}}{(1+d)^k} \cdot w_{ab}^d \longrightarrow \min \tag{8.15a}$$

$$\text{s.t.} \quad \sum_{c\in\mathcal{C}} f_a^c = 1 \qquad a \in \mathcal{A} \tag{8.15b}$$

$$f_a^{c_1} = \sum_{c_2\in\mathcal{C}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, c_1 \in \mathcal{C} \tag{8.15c}$$

$$f_b^{c_2} = \sum_{c_1\in\mathcal{C}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, c_2 \in \mathcal{C} \tag{8.15d}$$

$$w_{ab}^d = \sum_{\substack{c_1,c_2\in\mathcal{C}:\\ |c_1-c_2|=d}} s_{ab}^{c_1c_2} \qquad (a,b) \in \mathcal{A}^2, d \in \mathcal{D} \tag{8.15e}$$

$$f_a^c \in \{0,1\} \qquad a \in \mathcal{A}, c \in \mathcal{C} \tag{8.15f}$$

$$w_{ab}^d \in \{0,1\} \qquad (a,b) \in \mathcal{A}^2, d \in \mathcal{D} \tag{8.15g}$$

$$s_{ab}^{c_1c_2} \in \{0,1\} \qquad (a,b) \in \mathcal{A}^2, c_1, c_2 \in \mathcal{C} \tag{8.15h}$$

In W2-CP-CA, constraints (8.15e) are equivalent to constraints (8.3g) and (8.3h) in W2-CP by definition of the w-variables. Thus, W2-CP-CA can be viewed as an alternative formulation for W1-WCO discussed in Section 7.2.2. Furthermore, by Theorem 8.4, the LP relaxation of W2-CP-CA is at least as strong as that of W1-WCO.

Theorem 8.4. *The LP relaxation of W2-CP-CA is at least as strong as that of W1-WCO. Moreover, there are instances for which the LP bound provided by W2-CP-CA is strictly better than that provided by W1-WCO.*

Proof. Consider an LP-relaxation solution to W1-WCO where $f_a^c = \frac{1}{|\mathcal{C}|}, c \in \mathcal{C}$, and $w_{ab}^d = 0, (a,b) \in \mathcal{A}^2, d \in \mathcal{D}$. The solution is feasible since constraints (7.4b) and (7.4c) are satisfied, and all variables are non-negative. Observe that this LP solution is also optimal to the LP relaxation of W1-WCO because it gives zero as the objective function value, which is thus the LP bound provided by W1-WCO.

Because of the non-negativity requirement for variables in the LP-relaxation, the optimal LP-relaxation solution to W2-CP-CA is at least zero and therefore cannot be weaker than the LP bound provided by W1-WCO. Moreover, if all the v-parameters in (8.15a) are positive, the LP bound to W2-CP-CA is non-zero because there must be at least one non-zero f-variable for each AP (due to (8.15c)), which enforces some s-variables to be non-zero. This, in turn, results in that some w-variables are also non-zero, which gives an objective function value strictly greater than zero. Hence the conclusion. □

The channel assignment problems are very difficult to solve to optimality in general, and especially for large networks. Therefore, for finding a lower bound for problem W2-CP, it may be sufficient with a lower bound for W2-CP-CA. An overview of lower bounding techniques for channel assignment problems can be found, for example, in [8, 47] and the references therein.

8.4 An Upper Bounding Algorithm Based on Problem Decomposition

In Section 8.3, there has been described a way of decomposing W2-CP into the power assignment part and the channel assignment part. The idea has been exploited in the third

lower bounding approach and can also be utilized to derive an upper bound. As previously mentioned, a channel assignment obtained by solving W2-CP-CA is feasible in W2-CP. Combined with a feasible power assignment (i.e., satisfying constraints (8.3b), (8.3c), and (8.3j)), it gives a feasible solution to W2-CP and thus an upper bound for W2-CP. A feasible power assignment can be found, for example, by solving an optimization problem that involves constraints (8.3b), (8.3c), and (8.3j), and minimizes the total power consumption in the network, i.e.,

$$\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{L}} l\cdot p_a^l \ . \tag{8.16}$$

Given a feasible solution in the w-variables to W2-CP-CA and a feasible power assignment $\{l_a, a\in\mathcal{A}\}$, we can construct a feasible solution to W2-CP-CA in the y-variables as follows,

$$y_{ab}^{ld} = \begin{cases} 1 & \text{if } w_{ab}=1 \text{ and } l=l_a,\\ 0 & \text{otherwise,}\end{cases}$$

$$y_{ba}^{ld} = \begin{cases} 1 & \text{if } w_{ab}=1 \text{ and } l=l_b,\\ 0 & \text{otherwise.}\end{cases}$$

Observe that the objective function of W2-CP can be computed directly, i.e., without constructing the solution in the y-variables:

$$\sum_{(a,b)\in\mathcal{A}^2}\sum_{d\in\mathcal{D}} \frac{\nu_{ab}^{l_a}+\nu_{ba}^{l_b}}{(1+d)^k}\cdot w_{ab}^d \ . \tag{8.17}$$

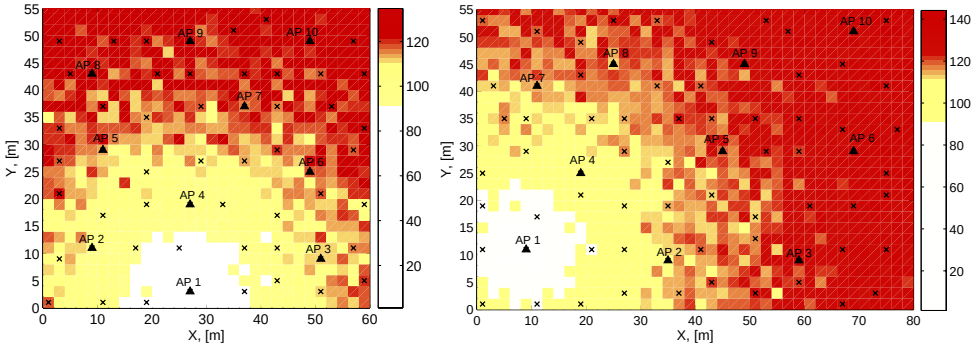
8.5 Numerical Results

In this section, we experiment with model W2-CP presented in Section 8.2. We solve the problem for two test networks considering five scenarios for each network. The networks and the scenarios are described in detail in Section 8.5.1. The reference configurations are presented in Section 8.5.2. For each test network and each scenario we find an optimal solution to problem W2-CP. The obtained solutions are then evaluated and compared to the reference configurations. The results are presented in Section 8.5.3. In Section 8.5.4 we experiment with the lower bounding approaches discussed in Section 8.3.

All numerical experiments have been conducted on HP ProLiant DL385 with two AMD Opteron Processor 285 (2.6 GHz, 1MB cache, 4GB RAM, dual-core). For finding optimal solutions to MIP and LP problems, we used commercial solver ILOG CPLEX [63]. The input data were translated into the required by solver format using modeling language ZIMPL [44], except for finding lower and upper bounds by decomposing the problem into the power assignment and the channel assignment subproblems. For the latter, we have run a C program with ILOG CPLEX Callable Library [63] that enables interaction with ILOG CPLEX MIP solver.

8.5.1 Test Networks and Test Scenarios

In the first test network, further referred to as the *dense network*, we place 10 APs in an area of 60 m × 55 m. In the second test network, referred to as the *sparse network*, the number of APs is 10, and the area size is 80 m × 55 m. The areas are represented as grids of square bins with resolution of 2 m×2 m. TPs are selected one per bin in the center of the bin. The number of TPs is 840 and 1120 for the dense network and the sparse network, respectively. Considering that many points can be covered by some AP with the minimum transmit power, the actual number of TPs that must be included in set $\mathcal{J}$ is 121 and 304 for the dense and the sparse network, respectively. In the sparse network, there are also a few bins (three, to be precise) that cannot be covered by any of the APs even with their maximum transmit power. These bins have also been excluded from set $\mathcal{J}$.



(a) Test network with dense AP placement

(b) Test network with sparse AP placement

Figure 8.4: Visualization of test networks.

Table 8.1: Test scenarios

Scenario name	Area [m×m]	Number of TPs in $\mathcal{J}$	Number of AP locations: $\lvert\mathcal{A}\rvert$ ($\lvert\mathcal{A}'\rvert$)	Number of channels $\lvert\mathcal{C}\rvert$	Channel set $\mathcal{C}$
			Dense network		
test_3d	60×55	121	10 (56)	3	$\{1, 6, 11\}$
test_4d	60×55	121	10 (56)	4	$\{1, 5, 9, 13\}$
test_5d	60×55	121	10 (56)	5	$\{1, 4, 7, 10, 13\}$
test_7d	60×55	121	10 (56)	7	$\{1, 3, 5, 7, 9, 11, 13\}$
test_6d	60×55	121	10 (56)	6	$\{1, 2, 3, 4, 5, 6\}$
			Sparse network		
test_3s	80×55	304	10 (70)	3	$\{1, 6, 11\}$
test_4s	80×55	304	10 (70)	4	$\{1, 5, 9, 13\}$
test_5s	80×55	304	10 (70)	5	$\{1, 4, 7, 10, 13\}$
test_7s	80×55	304	10 (70)	7	$\{1, 3, 5, 7, 9, 11, 13\}$
test_6s	80×55	304	10 (70)	6	$\{1, 2, 3, 4, 5, 6\}$

The AP locations have been chosen from a number of candidate locations randomly generated in the given areas (for each of the test networks) with uniform distribution. The AP placement was decided by solving an optimization problem that maximizes the total net user throughput in the network subject to a full coverage constraint and a constraint limiting the total number of installed APs by 10 (see model W1-NT in Section 7.2.1 for more details).

The transmit power levels of APs and APs have been selected in accordance with [16] and [18]. The MT transmit power is the same for all MTs and is fixed to 15 dBm. The set of available power levels at each AP is (in dBm) $\mathcal{L} = \{-1, 4, 7, 10, 13, 15\}$. Omni-directional antennas are assumed for all STAs in both test networks. Following the shadowing model, the path gain values have been computed as a function of distance, with a random shadowing component following normal distribution with zero mean and standard deviation of 6 dB (see Appendix C for more details on the path loss model). The serving threshold γ^{srv} and the CS threshold γ^{cs} are respectively set to -90 dBm and -110 dBm. The parameter setting used in our numerical experiments is summarized in Appendix C.

Figures 8.4(a) and 8.4(b) visualize the two test networks. The AP locations are marked with triangles, and the candidate locations are denoted with x-marks. In both figures, we use color to show the path loss values from AP 1 to any TP in the service area. Intentionally, the color scale is divided into three ranges. The white color denotes the serving range of AP 1 when its transmit power is -1 dBm (considering the antenna gain and the serving threshold,

this is the area where the path loss is below 91 dB). The yellow color is used to show the area where the received signal (with -1 dBm transmit power) is below the serving threshold but satisfies the CS threshold, i.e., the corresponding path loss values are in the range between 91 dB (exclusive) and 111 dB (inclusive).

From Figure 8.4(a), we observe that the first test network is not coverage limited, i.e., full coverage can be achieved at a relatively low transmit power (but not the minimum!) at the most of the installed APs. It will be shown later that a large number of possible power allocation solutions in this network significantly complicates the problem. Also, note that in this test network, for many APs, contention parameters are equal or very close to the size of the AP serving ranges. The number of possible power assignment solutions in the sparse network, on the other hand,is very restricted by the full coverage requirement (see Figure 8.4(b)). This should make it easier to solve the problem for the sparse network.

For each of the test networks, we consider five channel sets of various sizes. In the first four channel sets, the channels are equally spaced in the frequency spectrum and span over the entire spectrum. In the fifth channel set, we included six channels from the first half of the spectrum. A combination of a test network and a channel set is further referred to as a *test scenario*. Table 8.1 summarizes the statistics for the test scenarios we have considered for contention-aware optimization.

8.5.2 Reference Configurations

We have chosen two configuration strategies for defining our reference configurations. By the first strategy (strategy A), the transmit power of APs is fixed at its maximum level (15 dB). By the second strategy (strategy B), the transmit power of APs has been found by solving an optimization problem that minimizes the total AP transmit power in the network under the full coverage requirement (i.e., with the objective function (8.16) and constraints (8.3b), (8.3c), and (8.3j)). In both configuration strategies, the channel set consists of three channels $\mathcal{C} = \{1, 6, 11\}$, and the channel assignment is found by solving the minimum co-channel overlap model W1-CO presented in Section 7.2.2. The two configuration strategies have been applied in the two test networks (the dense network and the sparse network) discussed in Section 8.5.1.

The following notation is used to denote the resulting reference configurations,

- ref_Ad (strategy A, dense network),
- ref_Bd (strategy B, dense network),
- ref_As (strategy A, sparse network),
- ref_Bs (strategy B, sparse network).

8.5.3 Optimal Solutions

For evaluating and comparing different network configurations, we consider two measures: Contention span and DL interference. Note that when evaluating contention and interference, we assume that an MT is always assigned to the best AP (in terms of signal propagation), i.e., the one from which the strongest signal is received.

Contention span is defined for each TP j as an area from where contention can be expected, i.e., as a set of TPs where MTs may contend with the one located at TP j. We use this characteristic to evaluate contention for a given network configuration. Average contention span is computed as an average over all TPs.

The amount of interference is a straightforward measure since overlapping channels are involved. We consider only *DL interference*, i.e., the received signal power received from APs. Assume a TP is associated with a TP a operating on channel c. All signals received from other APs that use channels different from channel c are considered interfering. Signals received from APs operating on channel c are considered interfering only if the received signal

Table 8.2: Evaluation of reference configurations

Reference configuration	Average contention span*, [%]	Average DL interference, [dBm]	Objective function (8.3a)
ref_Ad	36.63	-90.70	6988.22
ref_Bd	33.65	-99.65	3445.92
ref_As	32.19	-94.16	6917.41
ref_Bs	30.94	-100.02	4296.96

* All four contention types are considered.

Table 8.3: Optimal integer and LP solutions obtained by contention-aware optimization

Scenario	Optimal solution				LP-relaxation solution		
	Objective function	CPU [sec]	Av. content. span*, [%]	Av. DL interference, [dBm]	Objective function	Gap [%]	CPU [sec]
test_3d	3061.47	59.92	34.96	-102.06	116.45	96.20	0.03
test_4d	2001.15	406.69	25.16	-99.08	77.67	96.12	0.04
test_5d	1468.83	2234.42	19.94	-95.33	75.90	94.83	0.04
test_7d	1001.94	17748.79	14.64	-93.58	74.59	92.56	0.04
test_6d	2143.25	7683.60	17.52	-88.71	351.06	83.62	0.04
test_3s	3764.63	6.11	33.22	-101.46	161.27	95.72	0.01
test_4s	2437.21	36.40	22.49	-94.89	104.73	95.70	0.01
test_5s	1818.78	231.86	17.97	-95.11	101.39	94.43	0.02
test_7s	1127.47	1090.18	14.01	-92.71	98.92	91.23	0.04
test_6s	2643.76	353.84	15.04	-88.14	466.07	82.37	0.03

* All four contention types are considered.

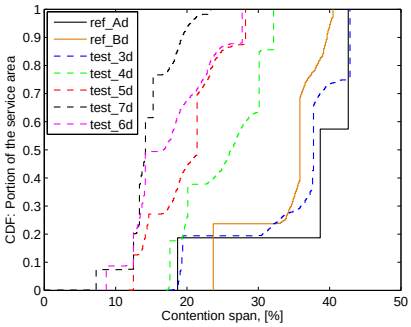
power is below the CS threshold (otherwise, contention occurs). To compute interference, we sum up (in linear scale) the received power of all interfering signals applying a scaling factor that depends on the channel distance. The factor is one when channel distance is zero. The scaling factors for channel distances $0 \leq d \leq 5$ are found as average of those for the same distance presented in [55]. For channel distances that are larger than five but at most eight, we define reasonably small factor values. Thus, we use factors of 1.0, 0.87, 0.75, 0.63, 0.31, 0.1, 0.05, 0.025, 0.01, for channel distances $0 \leq d \leq 8$, respectively, and zero for $d > 8$.

Table 8.2 presents the results of evaluating of the four reference configurations. Not surprisingly, the reference configurations with adjusted power yield better performance in terms of contention span and DL interference. In the last column of the table we also show computed values of objective function (8.3a).

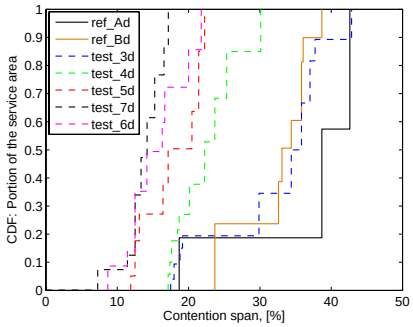
Table 8.3 presents optimal integer solutions, LP solutions, and computing times for all test scenarios. Additionally, we compute gaps to estimate the quality of the obtained lower bounds from which we observe that LP bounds for the studied model are very poor. Average contention span and DL interference metrics show a significant improvement over the reference configurations. Average figures are not very demonstrative, though. Therefore, we present CDF plots for contention span and interference for each of the test network. Exploring Figures 8.5(c) and 8.6(c), we find out that the best solutions from the interference point of view are the optimal solutions with three channels (i.e., for test_3d and test_3s) that perform better as compared to the corresponding reference configurations with adjusted power and significantly better than the corresponding reference configurations with the fixed power. The worst solutions are those obtained for the test scenarios with six channels.

Comparing Figure 8.5(a) to Figure 8.5(b) and Figure 8.6(a) to Figure 8.6(b), we observe that contention through APs form the most part of the total contention. This justifies our modeling approach that takes into account only contention of types (b) and (d).

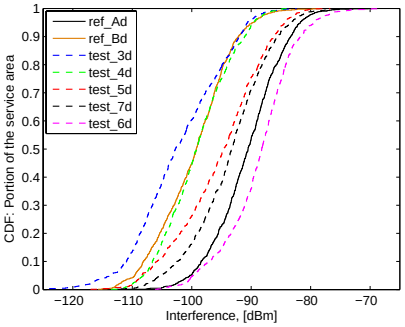
Table 8.4 presents channel assignment for the reference configurations and all the test scenarios. For each scenario/configuration, we show which APs operate on each of the channels.



(a) All four contention types

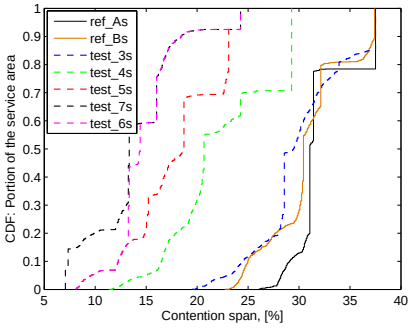


(b) Contention through an AP

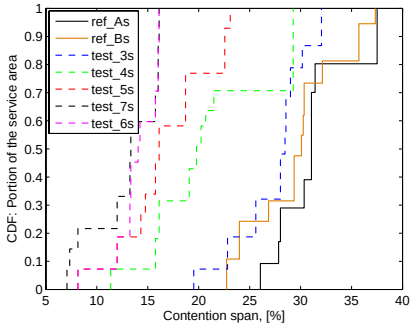


(c) DL interference

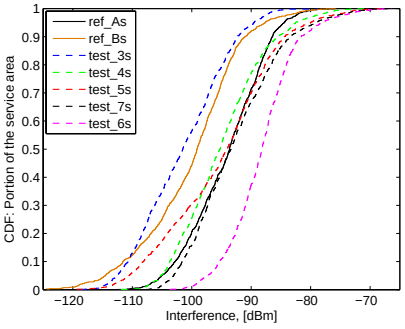
Figure 8.5: Contention span and DL interference CDFs for the dense test network.



(a) All four contention types



(b) Contention through an AP



(c) DL interference

Figure 8.6: Contention span and DL interference CDFs for the sparse test network.

Table 8.4: Channel assignment

Scenario	Channel c												
	1	2	3	4	5	6	7	8	9	10	11	12	13
ref_Ad	2;6;9	-	-	-	-	3;4;8;10	-	-	-	-	1;5;7	-	-
ref_Bd	2;3;9;10	-	-	-	-	4;6;8	-	-	-	-	1;5;7	-	-
ref_As	2;6;8	-	-	-	-	3;4;9	-	-	-	-	1;5;7;10	-	-
ref_Bs	2;6;7	-	-	-	-	1;5;8;10	-	-	-	-	3;4;9	-	-
test_3d	2;7;10	-	-	-	-	1;5;6;9	-	-	-	-	3;4;8	-	-
test_4d	5;6;7	-	-	-	1;4;9	-	-	-	3;8	-	-	-	2;10
test_5d	5;6	-	-	2;10	-	-	4;7	-	-	1;9	-	-	3;8
test_7d	2;10	-	5	-	6	-	8	-	3	-	1;9	-	4;7
test_6d	2;10	1;9	4;7	5	6	3;8	-	-	-	-	-	-	-
test_3s	6;7;8	-	-	-	-	3;4;5	-	-	-	-	1;2;9;10	-	-
test_4s	2;5;9	-	-	-	1;4;10	-	-	-	6;7	-	-	-	3;8
test_5s	6;7	-	-	5;8	-	-	2;9	-	-	1;10	-	-	3;4
test_7s	2;5	-	9	-	1;10	-	4	-	3	-	8	-	6;7
test_6s	6;7	8	3	2;5	1;10	4;9	-	-	-	-	-	-	-

Table 8.5: Channel coverage, in % to the total area $|\mathcal{J}|$

Scenario	Channel c												
	1	2	3	4	5	6	7	8	9	10	11	12	13
ref_Ad	38.69	-	-	-	-	42.62	-	-	-	-	18.69	-	-
ref_Bd	40.48	-	-	-	-	35.83	-	-	-	-	23.69	-	-
ref_As	31.34	-	-	-	-	30.98	-	-	-	-	37.41	-	-
ref_Bs	32.05	-	-	-	-	37.32	-	-	-	-	30.36	-	-
test_3d	33.74	-	-	-	-	19.40	-	-	-	-	42.86	-	-
test_4d	17.62	-	-	-	22.12	-	-	-	32.14	-	-	-	30.12
test_5d	21.43	-	-	23.21	-	-	12.50	-	-	14.64	-	-	28.21
test_7d	23.33	-	7.26	-	14.17	-	15.24	-	13.33	-	14.05	-	12.50
test_6d	22.86	14.05	12.50	8.69	14.16	27.74	-	-	-	-	-	-	-
test_3s	37.51	-	-	-	-	28.56	-	-	-	-	33.93	-	-
test_4s	20.68	-	-	-	25.78	-	-	-	24.26	-	-	-	29.27
test_5s	24.26	-	-	18.71	-	-	15.22	-	-	18.71	-	-	23.10
test_7s	13.34	-	7.34	-	18.71	-	7.07	-	16.03	-	13.25	-	24.26
test_6s	24.26	13.25	16.03	13.34	18.71	14.41	-	-	-	-	-	-	-

Studying the number of APs by channel, we observe that in the optimal channel assignment, more APs get assigned the most distant channels. Also, although the number of APs on each channel is not the same, it does not differ by more than one AP. Note that the result is compliant with Turán's theory. Similar observations are made even for sparse networks, where a different picture might be expected due to a stronger impact of the coverage requirement and therefore less flexibility in configuring the network. This, however, didn't happen because MTs always use a fixed (maximum) transmit power level, which results in larger CS ranges for MTs and strong effect of UL transmissions on the entire contention situation in the network.

Table 8.5 shows channel coverage statistics. For each scenario/configuration, we show the size of the area served on each particular channel, i.e., the channel coverage. From the table we observe that channel coverage is also quite balanced among the channels. Moreover, the most distant channels tend to have larger coverage in the network, which is reasonable. Interestingly, in configurations where the number of APs is larger on a channel in the middle of spectrum (test_3d) the total coverage area is even smaller than for the channels with a larger number of APs. This leads us to a conclusion that in an optimal configuration channel utilization is balanced among the channels.

Table 8.6: Lower bounds obtained by Approach 1, Approach 2, and their combination

Scenario	Approach 1			Approach 2			Approach 1 & 2		
	Obj.	Gap* [%]	CPU [sec]	Obj.	Gap* [%]	CPU [sec]	Obj.	Gap* [%]	CPU [sec]
test_3d	2088.59	31.78	0.10	317.29	89.64	0.52	2277.90	25.59	1.12
test_4d	1302.30	34.92	0.21	296.05	85.21	7.70	1462.24	26.93	3.33
test_5d	796.25	45.79	0.33	232.63	84.16	6.24	937.05	36.20	4.74
test_7d	469.94	53.10	0.70	232.57	76.79	32.06	619.01	38.22	25.02
test_6d	929.98	56.61	0.47	1021.20	52.35	35.96	1495.32	30.23	31.31
test_3s	2536.86	32.61	0.06	420.88	88.82	0.22	2978.33	20.89	0.34
test_4s	1396.97	42.68	0.08	380.05	84.41	1.45	1629.45	33.14	0.65
test_5s	767.51	57.80	0.11	304.89	83.24	1.77	1020.63	43.88	1.62
test_7s	409.10	63.72	0.23	302.39	73.18	9.51	620.22	44.99	5.05
test_6s	991.74	62.49	0.13	1327.03	49.81	12.16	1779.75	32.68	7.50

* Computed with respect to the optimal solutions in Table 8.3.

8.5.4 A Study on Lower Bounds

In this section, we study lower bounding solutions obtained by the three approaches presented in Section 8.3. Recall that lower bounds on the optimal objective function values for W2-CP computed from the LP relaxation have been already presented in Table 8.3. Observe that although this is probably the fastest way to compute a non-trivial lower bound for the studied problem (the computing times do not exceed 0.04 sec even for the largest channel set), the obtained lower bounds are very poor.

The effects of adding to the LP relaxation constraints (8.6) and (8.14) separately and together can be observed from numerical results presented in Table 8.6. Several observations have been made when exploring the results. First, constraints (8.6) (Approach 1) significantly strengthen the LP relaxation at a small increase in computing time. The computing times do not exceed 0.7 sec for any of the studied channel sets in both networks with the maximum time (0.7 sec) spent for the dense network with seven channels (test_7d). The best lower bounds (i.e., with the smallest gap) have been found when three channels are considered. This is reasonable (and expected) because the number of AP pairs using the same channel is the largest for the smallest channel set and thus, constraint (8.6) has the strongest influence on the LP bound. As another extreme, the constraint would have no effect at all if the number of APs and the size of the channel set would be equal (such an example is not considered among out test scenarios). Another observation is that Approach 2 is computationally more costly as compared to the Approach 1 and performs significantly worse when the channel set spans the entire spectrum. However, the lower bounds found by Approach 2 for scenarios using six channels from a half of the spectrum are better than those found by Approach 1. This is explained by that the objective function value is more sensitive to assigning channels that are neighbors in the considered (ordered) channel set to two neighboring APs due to a smaller channel distance than if the same number of channels is drawn from the entire spectrum.

Two interesting observations were made when constraint (8.6) and constraints (8.14) were simultaneously added to the LP relaxation of W2-CP. First, the computational times are smaller than when the Approach 2 is applied (except for the scenarios with three channels). Second, Approach 1 and Approach 2 strengthen different parts of the model which results in that the lower bounds obtained by applying them jointly are in most cases either larger or comparable to the sum of the lower bounds obtained by the two approaches separately.

Lower bounds obtained by Approach 3 are presented in Table 8.7. We observe that when the channel set is small (i.e., consists of three or four channels) computing times are smaller or comparable to those when we jointly apply Approach 1 and Approach 2. For larger channel sets, Approach 3 is more computationally expansive than Approach 1 and

Table 8.7: Lower and upper bounds obtained by decomposing the problem

Scenario	Lower bound (Approach 3)			Upper bound	
	Obj.	Gap* [%]	CPU [sec]	Obj.	Gap* [%]
test_3d	2362.36	22.83	0.12	3237.66	5.76
test_4d	1497.39	25.17	0.60	2246.30	12.25
test_5d	1071.88	27.02	3.97	1532.62	4.34
test_7d	787.88	21.36	47.79	1200.81	19.85
test_6d	1642.16	23.38	39.57	2331.95	8.80
test_3s	3447.36	8.43	0.16	3791.71	0.72
test_4s	2171.37	10.91	1.06	2499.02	2.54
test_5s	1625.96	10.60	6.68	1868.08	2.71
test_7s	1022.93	9.27	24.21	1223.91	8.55
test_6s	2363.20	10.61	13.50	2667.14	0.88

* Computed with respect to the optimal solutions in Table 8.3.

Approach 2 combined. Observe, however, that the lower bounds obtained by Approach 3 are far better than those obtained by any other approach considered here. Moreover, Approach 3 performs even better for the channel set sizes, i.e., when six or seven channels are considered (the most difficult test scenarios we study).

Except lower bounds, Table 8.7 presents also upper bounds obtained by reusing the channel assignment found by the last lower bounding approach and solving the minimum power problem for deciding on a power assignment (see Section 8.4 for more details). Solving the minimum power problem takes less than 0.01 sec. The time is thus spent mainly for the first part (finding a channel assignment). Comparing the upper bound solutions to the reference configurations, we find that the solutions are reasonably good, especially considering computing times.

To conclude, all the approaches proposed in Section 8.3 significantly outperform the LP relaxation results. Approach 2 have demonstrated the worst performance considering both the lower bound quality and the computing times. The advantage of the Approach 1 is that it allows us to get reasonably good lower bounds at a lower computational cost. Moreover, the obtained lower bounds can be further improved if Approach 1 is applied together with Approach 2. This, however, increasing computing times. The best lower bounds for all scenarios were obtained by Approach 3. The solution quality justifies the required computing times, especially for large channel sets. Relatively low computational costs and reasonably good lower bounds suggest that Approach 1, Approach 1 together with Approach 2, and Approach 3 can be effectively utilized in designing algorithms for solving problem W2-CP heuristically. Furthermore, combining them with the upper bounding approach, which is able to produce good solutions in a short amount of time, should facilitate finding near-optimal solutions of a good quality.

8.6 Conclusions

We have studied the problem of optimizing AP configuration in Wireless LANs considering channel assignment and AP transmit power as control parameters. We have presented a mathematical formulation of the problem that minimizes contention probability in the network subject to a coverage constraint. We have also proposed three approaches for lower bounding among which the approach based on strengthening the LP relaxation of the model and the approach based on problem decomposition have shown to be very efficient.

There has been performed an extensive numerical study with five channel sets of various sizes and with two networks characterized by different density of installed APs. The study showed that the proposed contention-aware model allows us to find a good balance between

contention and interference in the network. Compared to the reference configurations found by sequentially deciding on AP locations, power assignment, and then channel assignment, the optimal configurations significantly improve the interference and contention situation in the network. Moreover, the model has demonstrated the ability to provide reasonable network design solutions with respect to the considered channel set. The latter is particularly important for making the planning process more flexible in deciding on the channel sets at each AP. Another advantage of the model is that it can be tuned to change the priorities given to contention and interference (this is done by changing function $F(d)$).

Although we have been able to solve the problem to optimality for our test networks, it would be still interesting to develop a computationally efficient optimization approach that could be used for configuring large networks. An efficient heuristic algorithm would also be of great interest when extending the model with an option to also decide on AP locations (such a model extension is straightforward, but finding optimal solutions becomes then a really challenging task).

Bibliography

[1] IEEE Std 802.11–1997, Part 11: Wireless LAN Medium Access Control(MAC) and Physical Layer (PHY) specifications, 1997.

[2] ANSI/IEEE Std 802.11–1999 (R2003), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, 2003.

[3] IEEE Std 802.11a–1999 (R2003) (Supplement to ANSI/IEEE Std 802.11-1999(R2003)), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, High-speed Physical Layer in the 5 GHz band, 2003.

[4] IEEE Std 802.11b–1999 (Supplement to ANSI/IEEE Std 802.11-1999(R2003)), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Higher-speed Physical Layer extension in the 2.4 GHz band, 2003.

[5] IEEE Std 802.11F–2003, IEEE trial-use recommended practice formulti-vendor access point interoperability via an Inter-Access Point Protocol across distribution systems supporting IEEE 802.11 operation, 2003.

[6] IEEE Std 802.11g–2003 (Amendment to IEEE Std 802.11-1999 (R2003)), Part 11: Wireless LAN Medium Access Control (MAC) and PhysicalLayer (PHY) specifications, Further higher data rate extension, 2003.

[7] IEEE Std 802.11h–2003 (Amendment to IEEE Std 802.11-1999 (R2003)), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Amendment 5: Spectrum and transmit power management extensions in the 5 GHz band in Europe, 2003.

[8] K. Aardal, S. Van Hoesel, A. M. C. A. Koster, C. Mannino, and A.Sassano. Models and solution techniques for frequency assignment problems. *4OR: A Quarterly Journal of Operations Research*, 1(4):261–317, Dec. 2003.

[9] A. Adya, P. Bahl, R. Chandra, and L. Qiu. Architecture and techniques for diagnosing faults in IEEE 802.11 infrastructure networks. In *Proc. of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pages 30–40, Sep. 2004.

[10] M. Aigner. Turán's graph theorem. *American Mathematical Monthly*, 102(9):808–816, Nov. 1995.

[11] A. Akella, G. Judd, S. Seshan, and P. Steenkiste. Self-management in chaotic wireless deployments. *Proc. of the 11th Annual Intl. Conference on Mobile Computing and Networking (MobiCom '05)*, pages 185–199, Sep. 2005.

[12] E. Amaldi, S. Bosio, F. Malucelli, and D. Yuan. On a new class of set covering problems arising in WLAN design. In *Proc. of the Intl. Network Optimization Conference (INOC '05)*, volume B2, pages 470–478, Jan. 2005.

[13] E. Amaldi, A. Capone, M. Cesana, L. Fratta, and F. Malucelli. *Algorithms for WLAN coverage planning*, volume 3427 of *Lecture Notes in Computer Science (LNCS)*, chapter Mobile and Wireless Systems, pages 52–65. Springer, Feb. 2005.

[14] A. Caprara, M. Fischetti, and P. Toth. Algorithms for the set covering problem. *Annals of Operations Research*, 98:353–371, 2000.

[15] P. Chevillat, J. Jelitto, and H. L. Truong. Dynamic data rate and transmit power adjustment in IEEE 802.11 Wireless LANs. *International Journal of Wireless Information Networks*, 12(3):123–145, July 2005.

[16] Cisco Systems, Inc., `http://www.cisco.com/en/US/products/hw/wireless/ps430/`. *Cisco Aironet 1200 Series Access Points, Data sheet.*

[17] Cisco Systems, Inc, `http://www.cisco.com/warp/public/cc/pd/witc/ao350ap/`. *Cisco Aironet 350 Series Access Points, Data sheet.*

[18] Cisco Systems, Inc., `http://www.cisco.com/en/US/products/hw/wireless/ps4555/`. *Cisco Aironet 802.11a/b/g CardBus Wireless LAN Client Adapter, Data sheet.*

[19] Cisco Systems, Inc. Channel deployment issues for 2.4-GHz 802.11 WLANs. Technical report, `http://www.cisco.com/`, 2004.

[20] Cisco Systems, Inc. Using radio resource management to deliver secure and reliable WLAN services. White paper, July 2005.

[21] A. Eisenblätter. *Frequency assignment in GSM networks: Models, heuristics, and lower bounds.* PhD thesis, Technische Universität Berlin, Berlin, Germany, 2001.

[22] A. Eisenblätter, H.-F. Geerdes, and I. Siomina. Integrated access point placement and channel assignment for Wireless LANs in an indoor office environment. In *Proc. of the 8th IEEE Intl. Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*, June 2007.

[23] A. Eisenblätter, M. Grötschel, and A. M. C. A. Koster. Frequency planning and ramifications of coloring. *Discussiones Mathematicae Graph Theory*, 22(1):51–88, 2002.

[24] P. Fuxjäger, D. Valerio, and F. Ricciato. The myth of non-overlapping channels: Interference measurements in 802.11. In *Proc. of the Fourth Intl. Conference on Wireless On-demand Network Systems and Services (WONS 2007)*, Jan. 2007.

[25] E. Garcia Villegas, L. Faixó, R. Vidal, and J. Paradells. Inter-access point communications for distributed resource management in 802.11 networks. In *Proc. of the 4th ACM Intl. Workshop on Wireless Mobile Applications and Services on WLAN Hotspots (WMASH '06)*, pages 11–19, Sep. 2006.

[26] E. Garcia Villegas, R. Vidal Ferré, and J. P. Aspas. Implementation of a distributed dynamic channel assignment mechanism for IEEE 802.11 networks. In *Proc. of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2005)*, volume 3, pages 1458–1462, Sep. 2005.

[27] E. Garcia Villegas, R. Vidal Ferre, and J. Paradells Aspas. Load balancing in WLANs through IEEE 802.11k mechanisms. In *Proc. of the 11th IEEE Symposium on Computers and Communications (ISCC '06)*, pages 844–850, June 2006.

[28] E. Garcia Villegas, E. López-Aguilera, R. Vidal, and J. Paradells. Effect of adjacent-channel interference in IEEE 802.11 WLANs. In *Proc. of the 2nd Intl. Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom '07)*, Aug. 2007.

[29] J. K. Han, B. S. Park, Y. S. Choi, and H. K. Park. Genetic approach with a new representation for base station placement in mobile communications. In *Proc. of the 54th IEEE Vehicular Technology Conference (VTC2001-Fall)*, pages 2703–2707, Oct. 2001.

[30] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda. Performance anomaly of 802.11b. In *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, pages 836–843, San Francisco, CA, March 2003.

[31] A. Hills. Large-scale wireless LAN design. *IEEE Communications Magazine*, 39(11):98–107, Nov. 2001.

[32] A. Hills and B. Friday. Radio resource management in wireless LANs. *IEEE Radio Communications Magazine*, 42(12):S9–14, Dec. 2004.

[33] W. H. Ho and S. C. Liew. Distributed adaptive power control in IEEE 802.11 wireless networks. In *Proc. of the Third IEEE Intl. Conference on Mobile Adhoc and Sensor Sysetems (MASS 2006)*, pages 170–179, Oct. 2006.

[34] S. Hurley. Planning effective cellular mobile radio networks. *IEEE Transactions on Vehicular Technology*, 12(5):243–253, 2002.

[35] K. Jaffrès-Runser, J.-M. Gorce, and S. Ubéda. QoS constrained wireless LAN optimization within a multiobjective framework. *IEEE Wireless Communications*, 13(6):26–33, Dec. 2006.

[36] R. Jain, D. M. Chiu, and W. Hawe. A quantitative measure of fairness and discrimination for resource allocationin shared systems. DEC Research Report TR-301, Sep. 1984.

[37] J. Jemai, R. Piesiewicz, and T. Kürner. Calibration of an indoor radio propagation prediction model at 2.4 GHzby measurements of the IEEE 802.11b preamble. COST 273 TD, Duisburg, Germany, 2002.

[38] J. Jemai and U. Reimers. Channel modeling for in-home wireless networks. In *Proc. of IEEE Intl. Symposium on Consumer Electronics (ISCE02)*, pages F123–F129, Erfurt, Germany, Sep. 2002.

[39] M. Johansson. Stora testet: WLAN-simulering. *Nätverk & Kommunikation*, 2:30–38, Feb. 2007.

[40] M. Kamenetsky and M. Unbehaun. Coverage planning for outdoor wireless LAN. In *Proc. of Intl. Zurich Seminar on Broadband Communications, 2002. Access, Transmission, Networking (IZS)*, Zurich, Switzerland, Feb. 2002.

[41] A. Kamerman and G. Aben. Throughput performance of Wireless LANs operating at 2.4 and 5 GHz. In *Proc. of the 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2000)*, volume 1, pages 190–195, London, UK, Sep. 2000.

[42] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J.W. Tatcher, editors, *Complexity of Computer Computations*, pages 85–103. New York: Plenum, 1972.

[43] R. M. Karp. On the computational complexity of combinatorial problems. *Networks*, 5:45–68, 1975.

[44] T. Koch. *Rapid Mathematical Programming*. PhD thesis, TU Berlin, Germany, 2004. Available at `http://www.zib.de/Publications/abstracts/ZR-04-58/`, ZIMPLis available at `http://www.zib.de/koch/zimpl`.

[45] C. E. Koksal, H. Kassab, and H. Balakrishnan. An analysis of short-term fairness in Wireless Media Access protocols. In *Proc. of the 2000 ACM SIGMETRICS Intl. Conference on Measurement and Modeling of Computer Systems*, pages 118–119, June 2000.

[46] A. M. C. A. Koster. *Frequency Assignment – Models and Algorithms*. PhD thesis, Maastricht University, 1999.

[47] A. M. C. A. Koster, S. P. M. van Hoesel, and A. W. J. Kolen. Lower bounds for minimum interference frequency assignment. *Ricerca Operativa*, 30:101–116, 2000.

[48] M. Lacage, M. H. Manshaei, and T. Turletti. IEEE 802.11 rate adaptation: A practical approach. In *Proc. of the 7th ACM Intl. Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM 2004)*, pages 126–134, Oct. 2004.

[49] Y. Lee, K. Kim, and Y. Choi. Optimization of AP placement and channel assignment in wireless LANs. In *Proc. of the 27th Annual IEEE Conference on Local Computer Networks (LCN'02)*, 2002.

[50] K. K. Leung and B.-J. Kim. Frequency assignment for IEEE 802.11 wireless networks. In *Proc. of the 58th IEEE Vehicular Technology Conference (VTC2003-Fall)*, Orlando, FL, Oct. 2003.

[51] X. Ling and K. L. Yeung. Joint access point placement and channel assignment for 802.11 wireless LANs. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC 2005)*, New Orleans, LA, March 2005.

[52] J.-L. Lu and F. Valois. Performance evaluation of 802.11 WLAN in a real indoor environment. In *Proc. of the 2nd IEEE Intl. Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2006)*, pages 140–147, June 2006.

[53] P. Marinier and A. Cuffaro. Power control in 802.11 Wireless LANs. In *Proc. of the 64th IEEE Vehicular Technology Conference (VTC2006-Fall)*, Oct. 2006.

[54] K. Medepalli and F. A. Tobagi. On optimization of CSMA/CA based Wireless LANs: Part I – Impact of exponential backoff. In *Proc. of 2006 IEEE International Conference on Communications (ICC 2006)*, volume 5, pages 2089–2094, June 2006.

[55] A. Mishra, S. Banerjee, and W. Arbaugh. Weighted coloring based channel assignment for WLANs. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(3):19–31, 2005.

[56] A. Mishra, V. Shrivastava, S. Banerjee, and W. Arbaugh. Partially overlapped channels not considered harmful. In *Proc. of the joint Intl. Conference on Measurement and Modeling of ComputerSystems (SIGMETRICS '06)*, volume 34, pages 63–74, June 2006.

[57] J. Riihijärvi, M. Petrova, and P. Mähönen. Frequency allocation for WLANs using graph colouring techniques. In *Proc. of the Second Annual Conference on Wireless On-Demand Network Systemsand Services (WONS '05)*, St. Moritz, Switzerland, 2005.

[58] R. C. Rodrigues, G. R. Mateus, and A. A. F. Loureiro. On the design and capacity planning of a wireless local area network. In *Proc. of Network Operations and Management Symposium (NOMS 2000)*, Honolulu, Hawaii, 2000.

[59] K. U. Rommel. NETIO benchmark, version 1.26. `http://www.ars.de/ars/ars.nsf/docs/netio`, 1997–2005.

[60] H. D. Sherali, C. M. Pendyala, and T. S. Rappaport. Optimal location of transmitters for micro-cellular radio communication system design. *IEEE Journal on Selected Areas in Communications*, 14(4):662–673, May 1996.

[61] I. Siomina. Wireless LANs planning and optimization. STSM Technical Report, COST Action TIST 293, Dec. 2005.

[62] Open source software. *Introduction to* `lp_solve` *5.5.0.10.* `http://lpsolve.sourceforge.net/5.5/`, March 2007.

[63] ILOG, Inc. *ILOG CPLEX 10.0, User's manual*, Jan. 2006.

[64] P. Turán. On an extremal problem in graph theory (in Hungarian). *Matematikai és Fizikai Lapok*, 48:436–452, 1941.

[65] P. Wertz, M. Sauter, G. Wölfle, R. Hoppe, and F. M. Landstorfer. Automatic optimization algorithms for the planning of wireless local area networks. In *Proc. of the 60th IEEE Vehicular Technology Conference (VTC2004-Fall)*, Los Angeles, CA, Oct. 2004.

[66] L. Yang, P. Zerfos, and E. Sadot. Architecture taxonomy for control and provisioning of wireless Access Points (CAPWAP). RFC 4118, June 2005.

[67] F. A. Zdarsky, I. Martinovic, and J. B. Schmitt. On lower bounds for MAC layer contention in CSMA/CA-based wireless networks. In *Proc. of Workshop on Discrete Algothrithms and Methods for MOBILE Computing and Communications (DIALM'05)*, pages 8–16, Cologne, Germany, 2005.

[68] C. Zhu, T. Nadeem, and J. R. Agre. Enhancing 802.11 wireless networks with directional antenna and multiple receivers. In *Proc. of the 25th IEEE Military Communications Conference (MILCOM 2006)*, Oct. 2006.

Part III

Managing Dynamic Power-efficient Broadcast Topologies in Ad Hoc Networks

Chapter 9

Introduction to Ad Hoc Networks

Due to rapid growth in popularity of wireless networks and portable mobile devices, ad hoc wireless networks have received significant attention. Low cost and the ease of deployment and maintenance have made them particularly attractive in situations when creating a network with a fixed infrastructure is difficult and/or inefficient. On the other hand, ad hoc networking gives rise to new challenges one of which is designing protocols that would allow for power efficient communication and prolonging network lifetime in energy constrained environments. The two issues are especially important when developing broadcasting techniques since broadcast is a very frequent and resource demanding operation in dynamic ad hoc networks. In this context, we have designed two algorithmic frameworks one of which addresses the network lifetime issue in stationary networks, i.e., where nodes are not mobile, and the other one focuses on power efficiency in mobile networks. In this chapter we introduce the concept of ad hoc networking, discuss some research issues, and give a short overview of related work. The two algorithmic frameworks are presented in Chapter 10 and Chapter 11, respectively.

9.1 The Concept of Ad Hoc Networking and Its Application

In Latin, *ad hoc* literally means "for this" being further interpreted as "for this purpose only". Generally, the phrase is used to refer to a solution that has been custom designed for a specific problem, is non-generalizable and cannot be adapted to other purposes. *Ad hoc networking* is a technology that enables devices to establish a self-organizing communication network without the need of any pre-existing infrastructure. In an ad hoc network, each device may act as a network *node* that can combine functionality of a terminal and a forwarding device. Ad hoc devices can take different forms, e.g., laptop computers, PDAs (Personal Digital Assistants), mobile phone, sensors, etc.. The network topology in an ad hoc network is established on the fly, depending on the current situation (e.g., connectivity, distribution of remaining battery energy, running application, security aspects, etc.), and should allow for adding and removing network devices with the minimum disruption of functionality of the entire network. Ad hoc networking can thus be viewed as a paradigm by which networks can be established spontaneously without efforts and costs for building up and maintaining a network infrastructure.

Ad hoc networks are typically *dynamic* by nature since in general they are designed for self-organizing communication which implies that the networks must be self-reconfigurable. Ideally, this means that the underlying routing infrastructure must be changing over time. Note that dynamic behavior of a network does not imply that the network nodes are necessarily mobile. Thus, a dynamic ad hoc network can be *stationary* (i.e., when the nodes' positions are fixed over time) or *mobile* (when some or all nodes may change their positions). As a special case, an ad hoc network can be *static*, when, for example, the infrastructure after being established does not change [26]. There is, however, a lot of confusion in literature

between static and stationary ad hoc networks.

Mobile ad hoc networks, *wireless sensor networks*, and *wireless mesh networks* are the types of wireless ad hoc networks. Mobile ad hoc networks are ad hoc networks consisting of wireless mobile devices. IETF defines a mobile ad hoc network as an autonomous system of mobile routers (and associated hosts) connected by wireless links, the union of which form an arbitrary graph; the routers are free to move randomly and organize themselves arbitrarily, and the networks wireless topology therefore may change rapidly and unpredictably [15]. This type of ad hoc networks became popular when laptops and IEEE 802.11 wireless networks started to become widespread in the mid- to late 1990s (see Section 6.2 for more details on the ad hoc mode for infrastructureless IEEE 802.11 WLANs).

Wireless sensor network is a wireless network consisting of spatially distributed autonomous (mobile or stationary) devices using sensors to cooperatively monitor physical or environmental conditions, e.g., temperature, sound, pressure, motion, at different locations (see, for example, [2, 36, 39] and the references therein for more details). Typically, devices in a wireless sensor network are characterized as low-cost and low-power radio devices.

Wireless mesh networks consist of mesh routers and mesh clients, where mesh routers have minimal mobility and form the backbone of the networks. Mesh routers are usually equipped with multiple wireless interfaces built on either the same or different wireless access technologies. A good survey on wireless mesh networks can be found, for example, in [1].

The range of possible situations in which ad hoc networking can be exploited is huge. A robust ad hoc networking scheme frees the individual from the geographical constraints of the fixed network. In this respect, it is fundamentally different from established mobile networking, in which mobile nodes need to remain within the coverage of a wireless base station, connected to the fixed network infrastructure. Below are presented some applications of ad hoc networking.

- **Military:** The first project in infrastructureless networking (known as PRNet, or Packet Radio Network) was initiated in 1972 by DARPA (Defense Advanced Research Projects Agency) and demonstrated the technologies to create a mobile ad hoc network supporting mobile users from a collection of broadcast, spread-spectrum radios. In the military context, the ad hoc networks can be used, for example, for establishing communication among a group of soldiers for tactical operations when setting up of a fixed infrastructure in enemy territories or in inhospitable terrains may not be possible.

- **Emergency services:** A potentially huge public application of ad hoc networking is the area of emergency services. Fire-fighters, police and others sometimes have to operate in areas where no information infrastructure is present and operations still need to be coordinated. Another scenario would be a situation after a great natural disaster, atomic reactor melt down or other catastrophes where an existing infrastructure is destroyed. To be able to analyze, for example, a possible atomic reactor leak, it may be a better solution to place tiny sensors throughout the area, instead of risking human lives by sending them to hazardous areas.

- **Vehicular communication:** A vehicular ad hoc network is an application of mobile ad hoc networking enabling communications among nearby vehicles and between vehicles and nearby fixed equipment, usually described as roadside equipment.

- **Fairs and conferences:** Providing wireless communication in spontaneously appearing commercial zones (e.g., markets, conferences, fairs, festivals) is another example application of ad hoc networking.

- **Personal Area Networks (PANs):** Personal Area Networks are used in a smaller range than WLANs and are intended to operate in a personal environment, connecting mobile phones with each other, or with PDAs, laptops, etc.. There are a lot of scenarios of using these networks, alongside with other WLAN-based solutions.

- **Multimedia:** Modern mobile telecommunication devices have become real wonders of miniaturization. Almost any mobile phone has got an integrated camera, microphone, MP3 player, etc.. Other multimedia applications will surely be added in the recent future. With more and more advanced multimedia devices and applications, the need of multimedia information exchange between heterogeneous devices rapidly increases along with the amount of the transmitted information.

- **Monitoring:** Area monitoring is a typical application of wireless sensor networks. Deployed over a target area, wireless sensors can be used to monitor a process or particular physical characteristics (heat, pressure, sound, light, electro-magnetic field, vibration, etc.) and to detect events of interest. Some examples of the application are environmental monitoring, habitat monitoring, acoustic detection, seismic detection, inventory tracking, medical monitoring.

- **Automation:** Service robots will become more and more mobile in the next years. The ad hoc technology shall help coordinating their work, increasing efficiency and making them independent of a fixed communications environment.

- **Ubiquitous Computing:** Ubiquitous computing is the concept of making many computers available throughout the physical environment while making them effectively invisible to the user. The examples are smart offices, classrooms, and homes that allow computers to monitor what is happening in the environment. Since ad hoc networking works without infrastructures and is by nature more user friendly than fixed solutions, it plays an important role in the ubiquitous computing considerations.

9.2 Research Challenges in Ad Hoc Networking

The technological challenges of ad hoc networking are not trivial. They are present at all layers of the OSI (Open Systems Interconnection) model. At the physical layer, the wireless medium is subject to signal interference, and a finite communication range depends on the transmitter power. Unlike Ethernet, the link layer of a wireless network is not able to detect collisions during transmissions; therefore, reliability must be ensured via mechanisms such as direct acknowledgement of reception. The network (or routing) layer must adopt to a dynamic topology in which neighbor sets change, sometimes frequently, over time. Routing protocols must be fault tolerant to ensure robustness in link failures, i.e., if a link breaks, the protocol should be be able to quickly select another path to avoid disruption. Note also that a hierarchical routing structure, such as that of Internet, is difficult and expensive to maintain when links change frequently. At the transport layer, protocols must be able to differentiate between congestion and frequent packet loss in order to deliver a reliable stream of data. Ad hoc networks also need very specialized security methods that have to deal with link-level security, secure routing, key management, privacy, etc.. Ad hoc networking becomes even more challenging considering that the nodes in an ad hoc network are in general non-homogeneous which sometimes also puts very hard restrictions on the amount of hardware per node (for example, in wireless sensor networks).

Scalability is another important issue to be addressed when designing protocols for ad hoc networks. The number of nodes may be on the order of hundreds or thousands (for example, in sensor networks), and the nodes' density that together with the coverage range defines the nodes' degree (number of direct links with other nodes) may vary from a few to tens links per node. This may lead to significant communication overhead and low overall network performance, especially in the presence of mobility.

Ensuring power efficiency and prolonging network lifetime are the other challenging tasks in designing protocols for ad hoc networks. Power-efficient communication implies minimization of nodes' power consumption during the communication process. One of the reasons for

this is to save power at a node level as well as from a network perspective. In addition to this, the amount of power spent in a communication session directly relates to the interference level in the network and affects the degree of contention for the medium among different nodes as well as collision probability. This justifies the need of designing protocols ensuring *minimum-power communication*. This objective also leads to a longer battery life, i.e., prolonging *node lifetime*. However, it does not necessarily maximize the *network lifetime* which is also very important when devices are run on batteries and power supply is either impossible or very limited, especially in sparse networks where switching off a node may break network connectivity. Therefore, routing protocols designed to prolong network lifetime usually take also into account the residual battery energy (in addition to transmit power). Ensuring power efficiency in communication is usually referred to as *power management*.

Power management in ad hoc networks spans all layers of the protocol stack and utilizes the information available at each of the levels. This enables design of different both layer-specific and cross-layer power management mechanisms. For example, the MAC layer does power management using local information while the network layer can take a more global approach based on topology or traffic characteristics. Below are the most common strategies for power-efficient communication in ad hoc networks,

- turning off non-used transceivers/receivers to conserve the battery energy;
- scheduling competing nodes to avoid wastage of energy due to contention and collisions;
- reducing communication overhead, e.g., by avoiding redundant transmissions or deferring the transmission when the channel conditions are poor;
- using power control.

The first strategy usually associated with the network lifetime objective and implies transition between different modes rated by the amount of consumed power, e.g., the IEEE 802.11 MAC defines the transmitting, receiving, sleeping, and power-off modes. The other three strategies require the topological information and therefore involve the network layer power management mechanisms. The second strategy, however, is more applicable on an existing topology, whilst the other two can be utilized for constructing communication topologies. Any of the three strategies can be adapted to provide either communication extending network lifetime or minimum-power communication. Moreover, each of the three strategies can be effectively combined with the first strategy with which they overlap the least (an example of such a combination can be found in [52]).

Before we start discussing different approaches on how to resolve the aforementioned research issues in the context of power-efficient broadcast in ad hoc networks, we first provide a reader with more information on broadcast techniques in ad hoc networks in general which is the purpose of the next section.

9.3 Broadcast Techniques in Ad Hoc Networks

Broadcasting is a process of a distributing a message from a source node to all other nodes in the network (one-to-all communication). Broadcasting is a common operation in ad hoc networks which is used not only for distributing the data of broadcast nature among all network devices, but also for exchanging control information, e.g., to support routing, paging, etc. The broadcast operation is even more important in mobile networks where the network infrastructure may change rapidly and in an unpredictable manner.

A straightforward and the simplest broadcasting method is

- *blind flooding* by which each node retransmits a broadcast message whenever it receives the packet for the first time.

Although blind flooding may be very reliable for low-density and highly mobile ad hoc networks, this broadcast method is known to suffer in general from serious redundancy, contention, and high probability of collisions (referred to as the *broadcast storm problem* [34]). Furthermore, the method performs very poorly in energy and power efficiency which is especially important in energy-supply limited and interference-limited environments.

In addition to the blind flooding protocol and its variations, Williams and Camp [48] suggested three other categories of broadcasting methods:

- *probability based methods* where each node rebroadcasts with some probability),
- *area based methods* in which each node rebroadcasts a message received from a neighbor either within a certain distance or from a certain location,
- *methods based on neighbor knowledge* is a group of methods in which the forwarding decision is based on the neighborhood knowledge (e.g., different pruning schemes and CDS-based algorithms).

Probability based and area based methods are simple solutions for reducing message redundancy in the network, while methods based on neighbor knowledge usually are more complicated and typically involve more computations. Therefore, the protocols from the first two groups are typically *reactive*, or *on-demand*, protocols, i.e., requiring the forwarding decisions to be made immediately for each broadcast message. Due to complexity, many protocols in the third group are *pro-active* and are designed to dynamically maintain a broadcast infrastructure common for all broadcast messages. An example of such infrastructure is *virtual backbone* composed of a subset of devices used to relay broadcast messages (see, for example, [38]). Broadcast messages originated at any device are guaranteed to reach all other devices if two conditions hold. First, every non-backbone device has a direct connection to at least one device in the backbone. Second, devices in the backbone form a connected network. In graph terminology, such a backbone is called a *connected dominating set* (CDS). A set of nodes is a CDS if it forms a connected subgraph, every node in the network has at least one neighbor in the set, and the nodes in the set together form a connected graph, i.e., a graph in which there exists a path from any node to any other. Broadcast messages sent by any source are received by all nodes, if and only if the virtual backbone is a CDS.

Kouvatsos and Mkwawa [28] extended the classification of broadcast methods with two families of broadcast methods, namely,

- *cluster based methods* where some nodes, clusterheads, are responsible for propagating broadcast messages to all other members of the cluster, whilst other nodes, cluster gateways, are responsible for delivering the messages to other clusters (see, for example, [31]),
- *tree based methods* where a broadcast message at a node is forwarded only to the node's neighbors in a broadcast tree (here, it is also very common to distinguish between *source-dependent* and *source-independent* tree formation methods).

In both groups, methods utilize local information about neighbors (e.g., to elect a clusterhead or to construct tree branches) and therefore are related to the aforementioned group of methods based on neighbor knowledge. On the other hand, the methods also require some global knowledge about the network (e.g., to provide connectivity among clusterheads or to ensure that the constructed broadcast topology is a tree). Note that protocols used for clustering are most likely to be pro-active since this is a resource demanding operation. The way of constructing broadcast trees (on-demand or pro-actively), on the other hand, depends a lot on traffic intensity and traffic distribution. In highly mobile networks with very unbalanced traffic among nodes, constructing a source-dependent broadcast tree for each broadcast session may be inefficient. In this situation, managing a source-independent broadcast tree dynamically (pro-actively) is probably a better solution.

Stojmenović and Wu [41] characterized broadcast methods by the five criteria: determinism, network information, reliability, 'hello' message content, and broadcast message content. By the first criterion, a broadcast protocol can be either *probabilistic* or *deterministic*, depending on whether or not a random number selection was used to make decisions. The random number usage is limited to the network layer decision, i.e., the underlying MAC protocol may still use random back-off counters, for example, in a network layer deterministic protocol. By the second criterion, based on the amount of the state information, the authors distinguish *global* and *local* broadcast protocols, but at the same time point out that the two categories do not map directly to centralized and distributed protocols (e.g., centralized algorithms can also be applied in distributed setting, if a deciding node has full global information about the network). A global broadcast protocol, centralized or distributed, is based on global state information (see, for example, [22]). In local broadcasting, a distributed broadcast protocol is based on solely local state information. All protocols that utilize knowledge of only one- or two-hop neighborhood belong to this category. Except global and local protocols, there can be also intermediate categories, e.g., *quasi-global* and *quasi-local* as suggested by Wu and Lou [50]. In quasi-global broadcasting, a broadcast protocol is based on partial global state information (e.g., constructing a global spanning tree in a number of sequential propagations [5]). In quasi-local broadcasting, a distributed broadcast protocol is based on mainly local state information and occasional partial global state information (e.g., cluster networks where clusters are constructed locally for most of the time, whilst the chain reaction does occur occasionally). By the reliability criterion, broadcast protocols are subdivided into *reliable* or *unreliable* protocols, depending on whether every node in the network is reached or not (a broadcast message may be lost, for example, as a result of collision or because of bad network connectivity provided by the broadcast topology).

Addressing the design issues for ad hoc protocols discussed in Section 9.2, we can conclude that among the aforementioned broadcast techniques local broadcast protocols are of major interest to achieve good scalability. Among them, deterministic protocols have an advantage of being more reliable than probabilistic ones. Moreover, it is straightforward that the dynamic nature of ad hoc networks and limited energy resource necessitate the design of protocols that facilitate quick and resource-efficient adaption of the broadcast topology to changing conditions. With this in mind and also considering that broadcasting in ad hoc networks is a frequent operation, it becomes clear that the ability to construct and dynamically maintain a source-independent broadcast topology is a desirable property of a broadcast protocol in these networks.

In this section, we have discussed different broadcast techniques for ad hoc networks in general. In the next section, we present different research contributions to designing broadcast communication protocols with a major focus on power-efficiency. This is also the main scope of Part III.

9.4 Related Work

9.4.1 Designing a Virtual Backbone

We have mentioned in Section 9.3 that a broadcast topology represented by a virtual backbone can be viewed as a CDS in the context of graph theory. If transmit power power is not only constant but also uniform, minimizing the total power in a CDS-based broadcast session becomes equivalent to minimizing the backbone (CDS) size, i.e., finding a *minimum connected dominating set* (MCDS) which is a well-known $\mathcal{NP}$-hard optimization problem in graph theory [21]. Note that in this case, there is no need to distinguish between source-dependent and source-independent broadcasting. Literature on MCDS is extensive (see, for example, [9, 10, 16, 22, 29, 33, 40, 43, 49] and the references therein).

To prolong lifetime, a more appropriate model is weighted CDS (WCDS), in which the

weights of devices can be used to reflect residual energy. Methods for choosing dominators for WCDS (also referred to as clustering) can be found, for example, in [13, 14]. Guha and Khuller proposed centralized approximation algorithms for WCDS in [22, 23]. Basagni et al. [8] presented a performance comparison of several protocols for generating CDS or WCDS. Algorithms proposed in [7] and [32] also deal with WCDS. These algorithms allow for distributed implementations. To maximize network lifetime, solving WCDS once is clearly not sufficient. In [45], the authors proposed a distributed procedure for computing WCDS. The procedure is applied at some regular time interval to maximize lifetime.

Distributed backbone construction and management in the presence of node mobility have been addressed in a number of studies. However, the amount of literature dealing with mobility is much less extensive than that for stationary networks. Liang and Hass [30] proposed a scheme consisting of a distributed greedy algorithm to find a dominating set, and a link-state routing protocol for backbone formation in mobile ad hoc networks. In [4], Alzoubi et al. illustrated how node mobility can be dealt with by introducing various notification messages and distributed execution of the CDS algorithm in [3, 43]. More recently, Ju et al. [25] presented a fully distributed algorithm for dynamic backbone formation. In this algorithm, a node joins the backbone either to enable some neighbor(s) to access the backbone (i.e., to act as a dominator in the CDS), or to provide connectivity between neighbors (i.e., to act as a connector in the CDS), or both. In addition, the algorithm applies pruning to backbone nodes that are redundant. To account for mobility, a node checks, at regular time intervals, whether it should join or leave the backbone. Node priority is used to enable tie-breaking when there are several candidates for state change. The authors of [42] proposed a distributed scheme that applies simple rules to construct a partial backbone. If additional nodes are needed to provide connectivity, node selection can be carried out either deterministically (using node priority) or randomly. The scheme in [42] requires a fixed network topology during execution. (Repeated applications of the scheme can be used to handle mobility.) The schemes in [25, 42] impose that, if a pair of nodes are two hops away, then (at least) one of their common neighbors must be in the backbone. As a consequence, for any pair of nodes, the backbone must contain at least one shortest path in the underlying graph, where the path length is in hop count.

9.4.2 Designing a Power-controlled Broadcast Topology

The second approach for constructing a broadcast topology is to allow devices to adjust their transmit power [37]. When power adjustment is used as a means of topology control, all devices can relay a broadcast message (similar to blind flooding) but the devices can adjust the transmit power in order to preserve the battery energy and/or to less interfere with other nodes. The problem of finding an optimal set of transmit power levels in ad hoc networks is known as the *range assignment problem* [27]. Note that in a general range assignment problem the direct links between nodes can be either symmetric, or bidirectional, or unidirectional. Due to possible differences in transmit power levels at neighboring nodes, the broadcast algorithms based on adjustable transmit power give rise to another issue, a necessity to deal with unidirectional links that need a special treatment by MAC and network layer protocols. If the transmit power of all nodes is to be the same (which theoretically implies bidirectional links), the problem of unidirectional links disappears. Moreover, the range assignment problem gets simpler being reduced to finding the critical transmit range in the network.

Source-dependent minimum power broadcasting with adjustable transmit power has been studied extensively in the literature (see, for example, [6, 11, 19, 44, 46, 47]). However, minimum-power broadcasting does not maximize lifetime [12], partly because the optimization criterion is the total transmit power, which in most cases does not correspond to lifetime, and partly because the solution is a static broadcast tree per source. Das et al. [18] considered the problem of maximizing the expected lifetime of a broadcast tree. The authors showed

that this problem is polynomially solvable. In [17], the authors considered the multicast version of the problem, and presented an integer programming model. Kang and Poovedran [26] presented an extensive study of strategies for extending lifetime of broadcasting over a static or dynamic broadcast tree. They provided a theoretical analysis of a minimum spanning tree (MST) approach for prolonging the lifetime of broadcasting over a static tree, and discussed heuristics for constructing a dynamic tree. Floréen et al. [20] considered the complexity of maximizing lifetime of a multicast session, and proposed two centralized approaches. We note that the problems studied in [17, 18, 20, 26] are restricted to *single* source. Even with this restriction, maximizing lifetime on-line requires quite a lot of computational effort and message overhead [26].

Source-independent broadcasting with adjustable power has been studied to a rather limited extent in the literature. Papadimitriou and Georgiadis [35] presented an approximation algorithm for minimizing the total power of broadcasting over a single tree. Whether or not the algorithm can be extended to addressing lifetime remains open.

Chapter 10

Extending Network Lifetime in Stationary Networks

In this chapter we present an algorithmic framework that addresses the issue of maximizing network lifetime. Within the framework, we have developed two fully distributed algorithms for dynamic topology control in stationary ad hoc networks. The algorithms aim at maximizing network lifetime with respect to source-independent broadcasting.

The first algorithm has been designed to dynamically construct and maintain a virtual CDS-based backbone. We assume that transmit power is non-adjustable, but may vary by node. The algorithm is presented in Section 10.1.

The second algorithm adopts the topology control strategy based on adjusting transmit power levels of the network devices. The algorithm is presented in Chapter 10.2.

The contribution lies in the design of the two algorithms with the following features.

- The algorithms are fully distributed – they require neither global topology information nor centralized computation. Moreover, the computational operations in the algorithms are very simple.
- No location information is required by the algorithms. This feature is essential for applications where it is difficult to obtain location data due to the physical environment.
- The algorithms enable smooth topology control. The algorithms do not construct a new broadcast topology from scratch at regular time intervals. Instead, the update process involves only the devices that can restore connectivity, can be pruned (in the first algorithm), or adjust their transmit power (in the second algorithm).
- The control messages used by the algorithms are very short, and the message size is not dependent on network size.
- The amount of message overhead is not sensitive to the update interval but rather depends on network dynamics. The algorithms allow periodical update, but an update (which implies message overhead) is performed only if there exists some device with a critical energy level.
- Our algorithm design guarantees backbone connectivity throughout the update process. Backbone update is therefore transparent to broadcast traffic.

10.1 Distributed and Smooth Update of a Virtual Backbone

The algorithm presented in this section enables distributed dynamic update of a virtual backbone in stationary networks under the assumption of fixed transmit power and with focus on maximizing network lifetime. The algorithm dynamically adapts backbone to the

energy levels of devices. A device having a low energy level[1] is pruned from the backbone, provided that letting some other devices with better energy levels join the backbone can restore backbone connectivity. Although the idea itself is intuitive and straightforward, an algorithm implementing the idea must address a number of issues, such as detecting whether or not backbone update is necessary, finding devices that should join the backbone, as well as preserving backbone connectivity during the update process. A centralized approach can handle these issues easily, but it is not very practical in ad hoc networking.

The work differs from some other approaches for prolonging lifetime (e.g., [7, 14, 32, 45]) in a couple of aspects. First, the algorithm does not use two separate phases for clustering and backbone formation. Second, in an update the algorithm does not recompute a new backbone from scratch.

10.1.1 System Model

Model Definition

We model a wireless ad hoc network by an undirected connected graph $G = (V, E)$, where V and E denote the sets of devices (or nodes) and bidirectional links, respectively. The links are defined by nodes' transmit power, transceivers' and receivers' characteristics, and the propagation environment. If two nodes u and v are directly connected, there exists a link $(u, v) \in E$ and the two nodes are called *one-hop neighbors*. The set of one-hop neighbors of node u is denoted by $N(u)$. The set of *k-hop* neighbors of u ($k > 1$) is denoted by $N_k(u)$, i.e., $N_k(u)$ consists of nodes to which the shortest path distance from u, counted in hops, equals k (see Figure 10.1). Note that if node v is a two-hop neighbor of u, then the following relations hold: $v \notin N(u)$ and $N(u) \cap N(v) \neq \emptyset$.

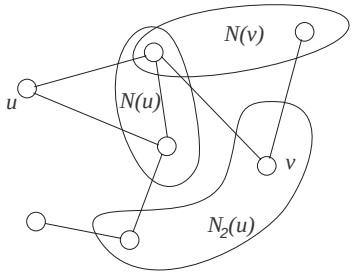


Figure 10.1: Neighborhood definition for an ad hoc network modeled as an undirected graph.

To model a one-hop transmission of a broadcast message in the network, we assume that all nodes utilize the *wireless multicast advantage* [46] by which a message transmitted by a node is received by all its one-hop neighbors. The concept of the wireless multicast advantage applies also to broadcast and implies that a multicast/broadcast transmission allows for resource saving (e.g., energy, bandwidth, and time) as compared to multiple one-hop transmissions.

A virtual backbone of the network consists of a subset of nodes (denoted by $D \subset V$). Only nodes in D (backbone nodes) can retransmit broadcast messages. Henceforth, we use the term *gateway* to refer to a backbone node. The network lifetime is defined as the time from the network initialization to the first node failure because of battery depletion.

[1] The meaning of low energy level will be defined more precisely in Section 10.1.2.

Connectivity Characterization

A broadcast message initiated at any node can reach all the other nodes if and only if D is a *connected backbone*, that is, a CDS. Connectivity of the network backbone can be characterized in terms of neighborhood connectivity [51]. As a result of this characterization, a node needs to monitor reachability between its neighbors only, i.e., connectivity information of other parts of the network is not required at the node. Interestingly, neighborhood connectivity is not only *necessary*, but also *sufficient* to define a CDS.

The condition, referred to as the *neighborhood connectivity condition*, is met if for every non-gateway node $u \in V \setminus D$, a broadcast message initiated at any of its neighbors $v \in N(u)$ is received by all other neighbors $\{w : w \in N(u), w \neq v\}$ through nodes in D. Observe that in the definition, node $w \in N(u)$ appears as either a direct neighbor of v or its two-hop neighbor. With this observation, the neighborhood connectivity condition is also met when a message transmitted by v is received by any node $w \in N_2(v)$ and trivially satisfied when $w \in N(v)$.

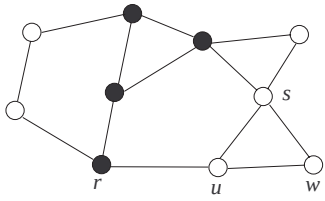


Figure 10.2: An example illustrating the neighborhood connectivity condition.

The condition is illustrated in Figure 10.2, in which backbone nodes are marked black. The condition is not satisfied at node u because messages of r do not reach w (and vise versa). The condition does not hold at s either. Letting u or s become gateway results in a connected backbone. Theorem 10.1 formalizes the sufficiency and necessity of the neighborhood-connectivity condition. Equivalently speaking, if neighborhood connectivity holds at every non-backbone node, then the backbone provides global network connectivity. (That neighborhood connectivity holds at backbone nodes is obvious.)

Theorem 10.1. *A set of nodes $D \subseteq V$ form a connected backbone, that is, a CDS, if and only if the neighborhood-connectivity condition is satisfied.*

Proof. See [51].

The neighborhood-connectivity condition enables distributed control of global connectivity. Utilizing this condition, a node is responsible for observing connectivity between its neighbors only.

10.1.2 Algorithm Description

Preliminaries

Let t denote time, and k ($k \geq 1$) to index time interval. The length of a time interval is denoted by Δ. Let the starting time be zero. Following these notation, the first time interval is $[0, \Delta)$, and the kth interval is $[(k-1)\Delta, k\Delta)$. We use t_0^k to denote the starting time of interval k, i.e., $t_0^k = (k-1)\Delta$. The operations performed by the algorithm are organized in cycles. A cycle consists of one or several steps, depending on whether the backbone is to be updated or not. One algorithm cycle is applied at the beginning of every time interval, that is, at time t_0^1, t_0^2, and so on. (We assume that the time needed to complete a cycle is much shorter than Δ.) D^k is used to denote the backbone (i.e., the set of gateways) after

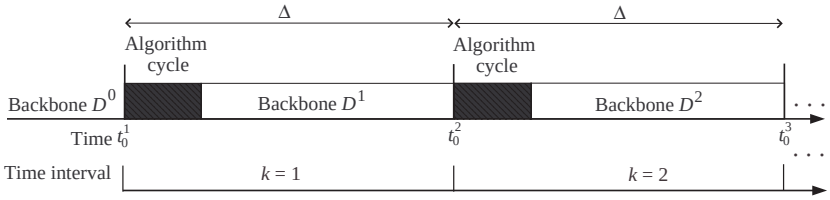


Figure 10.3: Time interval and algorithm cycle.

the completion of an algorithm cycle in time interval k. Further, we assume that the initial backbone at time t_0^1 is connected. This backbone is denoted by D^0. To construct D^0, any algorithm for constructing a CDS can be used (e.g., [7, 10, 16, 22, 29, 32, 33, 40, 43, 45]) as well as a slight modification of the algorithm. The notation introduced so far is illustrated in Figure 10.3.

The General Idea

Let us introduce the notion of *active* and *passive* gateways. An active gateway (AGW) relays all broadcast messages. A passive gateway (PGW), on the other hand, relays broadcast data messages, but does not relay all algorithm control messages. Informally speaking, a PGW tries to hide itself from the rest of the backbone, and "pretends" to be a non-gateway node. By doing so, some other nodes will believe that their neighborhood-connectivity condition does not hold, and consequently join the backbone in order to "restore" connectivity. (Note that for data messages, the backbone consists of AGWs as well as PGWs.) Which nodes that will join the backbone (i.e., become AGWs) is determined by their battery residual energy. When sufficiently many nodes have joined the backbone, a PGW will detect that its neighborhood-connectivity condition holds. At this stage, the PGW quietly prunes itself from the backbone and becomes a non-gateway node. If, however, a PGW cannot be replaced by nodes having better energy levels, it will go back to the active state.

A small example of backbone update is shown in Figure 10.4. In Figure 10.4(a), the black nodes form a connected backbone. Assume that v becomes passive (marked by the grey color in Figure 10.4(b)), and stops relaying control messages. Node u believes that the

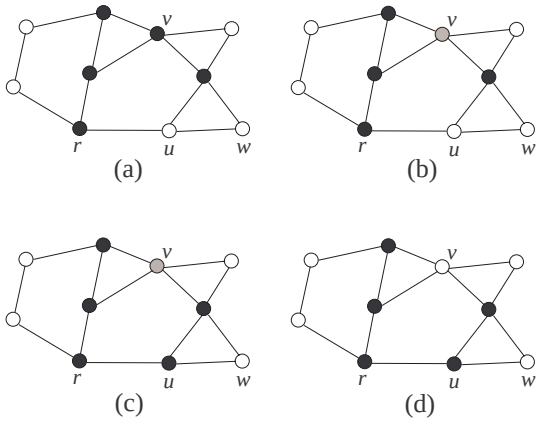


Figure 10.4: Backbone update: a small example.

Algorithm III.1 Managing dynamic virtual backbone in a stationary network

Input: $V, \Delta, D^0, t_0, t_{max}$

1: $k \Leftarrow 1$
2: $\mathcal{A} \Leftarrow D^0$ // Set of AGWs
3: $\mathcal{P} \Leftarrow \emptyset$ // Set of PGWs
4: **repeat**
5: $\quad \mathcal{C} \Leftarrow \emptyset$ // Set of candidate gateways
6: $\quad \mathcal{T} \Leftarrow \emptyset$ // Set of TPGWs
7: $\quad t_0^k \Leftarrow t_0 + (k-1)\Delta$
8: $\quad$ **wait until** $(t == t_0^k)$
9: $\quad \mathcal{T} \Leftarrow$ **Step1a**$(\mathcal{A}, \mathcal{P})$ // Select TPGWs
10: $\quad$ **if** $\mathcal{T} \neq \emptyset$ **then**
11: $\quad\quad \{\mathcal{A}, \mathcal{P}\} \overset{update}{\Leftarrow}$ **Step1b**$(\mathcal{T})$ // PGW election
12: $\quad\quad$ **Step2**(V) // Connectivity probing
13: $\quad\quad \{\mathcal{A}, \mathcal{C}, \mathcal{P}\} \overset{update}{\Leftarrow}$ **Step3**$(V \setminus \mathcal{A})$ // Connectivity control and pruning
14: $\quad\quad \{\mathcal{A}, \mathcal{P}\} \overset{update}{\Leftarrow}$ **Step4**$(\mathcal{C})$ // AGW election
15: $\quad\quad D^k \Leftarrow \mathcal{A} \cup \mathcal{P}$
16: $\quad$ **else**
17: $\quad\quad D^k \Leftarrow D^{k-1}$
18: $\quad$ **end if**
19: $\quad k \Leftarrow k + 1$
20: **until** $t > t_{max}$

neighborhood-connectivity condition does not hold as a control message of r does not reach w. Node u classifies itself as a candidate gateway. Since the condition does not hold at v either, v is also a candidate gateway. If u has a better energy level than that of v, u becomes an AGW, as shown in Figure 10.4(c). Now the condition becomes satisfied at node v, which prunes itself from the backbone, giving the result in Figure 10.4(d).

An AGW wishes to enter the passive mode if its residual energy level is low. Let $P_u(t)$ denote the level of residual energy of u at time t. The algorithm uses a reference level, denoted by P_u, which is the energy level associated with the most-recent time at which u became an AGW. The energy level of an AGW u is considered as low if $P_u(t) \leq \alpha P_u$, where α is a parameter chosen such that $0 < \alpha < 1$.

An AGW whose energy level has dropped below the threshold is further referred to as a *tentative passive gateway* (TPGW). There may be several TPGWs at a time, but at most one of them is allowed to become passive in an algorithm cycle.

The algorithm manages the backbone smoothly. In one cycle, updating the backbone involves some (usually very small) part of the network. Also, before the energy level of some AGW has dropped below its locally-defined threshold, there is no backbone update and hence zero message overhead. Moreover, as will be formally proved later in this section, pruning PGWs preserves backbone connectivity.

In the algorithm presentation, we make the assumption that nodes have knowledge of not only their neighbors, but also their two-hop neighbors. A node does not, however, know whether or not its broadcast messages can be received by a two-hop neighbor. The assumption is not crucial for the execution of the algorithm. It does, however, reduce the amount of memory requirement at nodes.

Algorithm III.1 is a sketch of the virtual backbone update algorithm presented in pseudo code. Note that the algorithm uses two time parameters, t_0 and t_{max}. Time t_0 is the time when the algorithm is to be started (backbone D^k is supposed to be formed by that time), and t_{max} is the end time point of the algorithm (can be infinite; note, however, that although the algorithm can perform even after the network becomes disconnected, it will stop in

those parts of the network where all nodes are connected through direct links or in a trivial network of a single node). The next algorithm cycle k is triggered at every node when the time t returned by its synchronized network timer is equal to t_0^k. In Algorithm III.1, each step of the algorithm is represented as a function of one or more sets of nodes that make a decision in the step. The result of each such function are the corresponding updated sets. For example, line 14 states that in Step 4 the decision on state transition is made by candidate nodes (i.e., nodes from set $\mathcal{C}$) and as a result, sets $\mathcal{A}$ and $\mathcal{P}$ get updated, i.e., the nodes may change their state to AGW or PGW. Note that we use $\Leftarrow$ to denote an assignment operation and $\overset{update}{\Leftarrow}$ for a set update operation.

Next we present in more detail each of the four steps involved in an algorithm cycle and give a description of a possible implementation of the algorithm.

An Algorithm Cycle

By this point, there has been presented the general idea of the algorithm. Let us now examine the implementation of the kth cycle of the algorithm. Recall that a cycle of the algorithm is composed by a sequence of four steps (some of them can be skipped under certain conditions). All the steps are outlined in Table 10.1.

Consider the first step of the algorithm cycle preformed at node u. At the beginning of time interval k (that is, at time t_0^k) the backbone D^k equals D^{k-1}. It has been previously mentioned that there may be selected at most one new PGW in each algorithm cycle. In addition to that, at t_0^k, there may be also some other PGWs that entered the passive mode in some earlier algorithm cycles and have not yet been pruned. The purpose of the first step is to determine PGWs through an election process.

Remark 1. If there is a PGW having lower energy level than all TPGWs, then none of the TPGWs will enter the passive mode. Otherwise the one with the lowest residual energy becomes passive.

Remark 2. If there is no PGW or TPGW at time $t = t_0^k$, then the algorithm cycle stops after Step 1. In this case, no additional energy or computational overhead is incurred. If we shorten Δ, which creates more potential update occasions, the additional algorithm overhead scales up less than proportionally, since only some of the extra occasions will be utilized.

Remark 3. The message **TentativePassive**(v, $P_v(t_0^k)$) sent by a PGW v has a couple of purposes. First, sending the message prevents a TPGW u from becoming passive if u has more residual energy than v. Second, in case there is no TPGW, this message will ensure that nodes do proceed to the next step of the algorithm cycle.

In the second step (see Step 2 in Table 10.1), node u uses a list, denoted by $L(u)$, to keep track on its direct neighbors and the two-hop neighbors from which control messages can be received. At the beginning of the step $L(u) = \emptyset, \forall u \in V$. Note that a **ConnectivityProbe** message is relayed by AGWs *only*, and consequently may or may not reach all nodes. Lists created in Step 2 are exchanged between neighbors in the next step.

In Step 3 (see Step 3 in Table 10.1), non-gateway nodes and PGWs care about **NodeList** messages from their neighbors, because these nodes are subject to state change. A PGW u is pruned if the neighborhood-connectivity condition holds at u. Note that in this case u knows about the existence of connectivity between its neighbors, but not through which gateways they are connected. Parameter K limits the number of consecutive cycles in which a gateway stays in the passive mode. As a result, the number of PGWs is at most K in any algorithm cycle. This limit is useful to avoid a scenario where many PGWs block each

Table 10.1: The key points of the algorithm cycle

Step 1 (Passive gateway election)
• An active gateway u checks its residual energy $P_u(t_0^k)$. If $P_u(t_0^k) \leq \alpha P_u$, u marks itself as a tentative passive gateway, and broadcasts message **TentativePassive**(u, $P_u(t_0^k)$) through backbone D^k. • A passive gateway v broadcasts message **TentativePassive**(v, $P_v(t_0^k)$) through backbone D^k. • If a tentative passive gateway u does not receive any **TentativePassive** message containing a lower energy value than its own, u becomes passive, otherwise u becomes active and sets $P_u = P_u(t_0^k)$. A tie can be broken using node identification.
Step 2 (Connectivity probing)
• Every node transmits a control message containing its identification. The message sent by v is denoted by **ConnectivityProbe**(v). • When u receives **ConnectivityProbe**(v), u does the following. 1. Node u checks whether $v \notin L(u)$ and $v \in N(u) \cup N_2(u)$. If both conditions hold, u adds v to $L(u)$. 2. If u is an AGW and this is the first time when message **ConnectivityProbe**(v) has been received in the algorithm cycle, u retransmits the message.
Step 3 (Connectivity control and pruning)
• Every node sends its list to neighbors. The message sent by v is denoted by **NodeList**(v, $L(v)$). • Non-gateway nodes and PGWs store **NodeList** messages. • A non-gateway node u checks whether $\forall v, w \in N(u)$, $w \in L(v)$. If not, u marks itself as a candidate gateway. • A PGW u checks whether $\forall v, w \in N(u)$, $w \in L(v)$. If so, u is pruned from the backbone and becomes a non-gateway node (i.e., $D^k = D^k \setminus \{u\}$). If u cannot be pruned and it has been a PGW in K consecutive algorithm cycles, u changes its state to active and updates P_u, otherwise u marks itself as a candidate gateway.
Step 4 (Active gateway election)
• A candidate gateway v broadcasts message **CandidateGateway**(v, $P_v(t_0^k)$) through backbone D^k. • If a candidate gateway u does not receive any **CandidateGateway** message having a higher energy value than its own, u becomes an active gateway (i.e., $D^k = D^k \cup \{u\}$) and updates P_u, otherwise u enters the state that u had in Step 3. A tie can be broken using node identification.

others' chance of being pruned. PGWs that become neither pruned nor active due to K are candidate gateways. Non-gateway nodes at which the neighborhood-connectivity condition does not hold are also candidate gateways.

In Step 4 (see Step 4 in Table 10.1), one of the candidate gateways becomes active. Note that, if there is no candidate gateway in Step 3, then no message is sent in Step 4. Also, pruning of PGW (in Step 3) and election of new AGW (in Step 4) may both take place in one algorithm cycle.

Figure 10.5 depicts the state transition diagram in an algorithm cycle. Note that the names of the transit states, i.e., in which nodes can only be within the algorithm cycle, appear in dashed frames. The states that nodes can have by the end of the time interval are shown in solid-line frames.

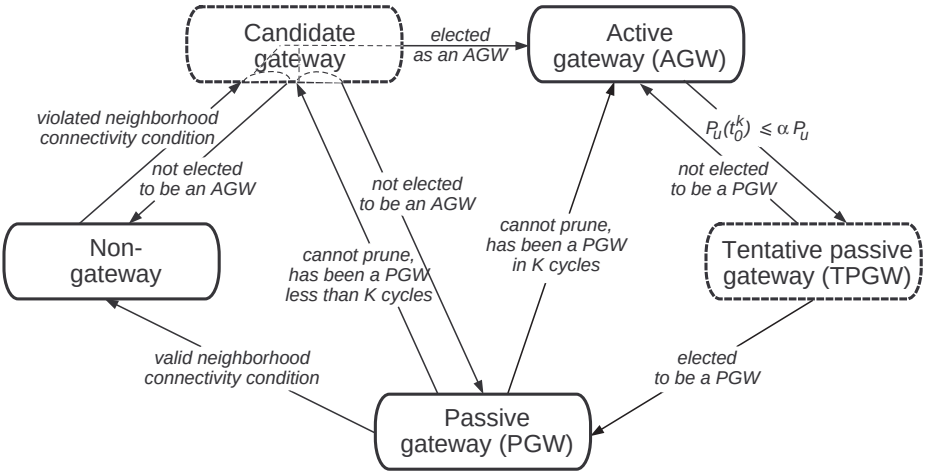


Figure 10.5: A state transition diagram for node u.

Backbone Connectivity

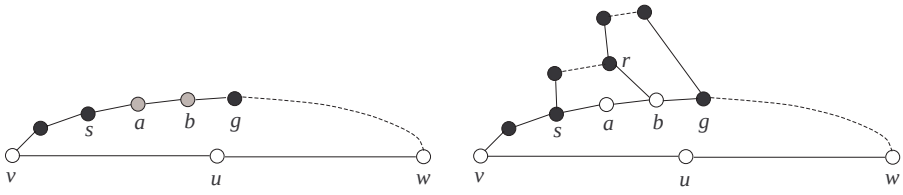
As was mentioned earlier, AGWs and PGWs together always form a connected backbone for broadcast data messages. Below we formalize this result and prove its correctness.

Theorem 10.2. *If D^0 is connected, then AGWs and PGWs together form a connected backbone during the entire network lifetime.*

Proof. Consider the backbone in the beginning of time interval k, i.e., at time t_0^k. At this time, the backbone is that from the previous step, i.e., D^{k-1}. Assume that D^{k-1} is connected. We show that backbone D^k in the end of the algorithm cycle is also connected. Obviously, the backbone connectivity may be broken only when a node prunes from the backbone, i.e., changes its gateway state to the non-gateway state. This may occur only in Step 3, which we further consider to examine whether the backbone remains connected after this step. Note, however, that at most one AGW may become PGW after Step 1, but the set of backbone nodes remains the same.

Consider the neighborhood connectivity condition (for broadcast data message) after Step 3. The condition holds obviously at AGWs and PGWs because they both relay data messages. Consider therefore a non-gateway node, say u, by the end of Step 3. If u was a PGW at the beginning of the step, then the condition is satisfied (and there exists a path of AGWs that did not change their status to PGW in Step 1) since otherwise u would not have been pruned. So the only case left is $u \notin D^{k-1}$ (i.e., u was a non-gateway node). Consider two neighbors, v and w, of u. After Step 1, the backbone must contain at least one path consisting of AGWs and PGWs from D^{k-1} connecting v and w. Figure 10.6(a) illustrates such a path containing AGWs as well as PGWs.

After Step 3, some of the PGWs in the path may have become non-gateway nodes, such as a and b. We start from v, and follow the path until we encounter the first non-gateway node a (see Figure 10.6(b)). Because a was pruned, at least one path of AGWs exists between its neighbors, in our case s and b. Denote the last AGW in the path by r. Because b became a non-gateway node, we repeat the argument, and conclude the existence of a path of AGWs between r and g. In the example g is an AGW, so we start from g and continue following the original path connecting v and w. Note that the construction is the same if g is a PGW. It can be easily realized that if we apply the procedure of finding an alternative path whenever necessary, we will eventually reach w. Therefore the condition is satisfied at u. We have



(a) A path with active and passive gateways

(b) Discovering alternative paths with active gateways

Figure 10.6: An illustration of backbone connectivity.

shown that the backbone connectivity is preserved when time evolves from t_0^k to t_0^{k+1}. This holds for any $k \geq 1$ and is also true for time interval $[t_0^{k+1}, t_0^{k+2}]$, which means that the backbone always remains connected starting from t_0^1. Hence the result of the theorem. □

Message Overhead

Assume that all four steps are executed in time interval k, and let $\bar{D}^k$ denote the set of AGWs after Step 1. We know $\bar{D}^k \subset D^k$. Steps 1, 2, and 4 involve broadcasting messages **TentativePassive**, **ConnectivityProbe**, and **CandidateGateway**, respectively, through either D^k or $\bar{D}^k$. In Step 2, the number of messages equals $|V|$. Because typically the numbers of nodes initiating **TentativePassive** messages (PGWs and TPGWs) and **CandidateGateway** messages (candidate gateways) are much less than $|V|$, message overhead in Step 2 is dominating among Steps 1, 2, and 4. Moreover, since Step 3 requires only one transmission per node, the amount of overhead in this step is also much less than that of Step 2.

In Step 2, the total number of transmissions is $|V||\bar{D}^k| < |V||D^k|$. The latter is the number of transmissions required to broadcast one message from every node to the rest of the network. Therefore, the number of transmissions in Step 2 is less than that for broadcasting $|V|$ data messages. Note that a control message is typically much shorter than a data message (indeed, message **ConnectivityProbe** contains essentially nothing else than node identification), and hence takes less transmit power.

Whether and how much lifetime can be prolonged using the algorithm depends on, in addition to message overhead, many other factors, such as network density and intensity of broadcast (data) traffic. In some cases the lifetime may even be shortened due to the algorithm. (This occurs, clearly, if there is a node that is the only possible connection between some other nodes.) In the next section we use simulation to quantify how lifetime is related to network characteristics as well as control parameters in the algorithm.

10.1.3 Performance Simulation

This section presents performance simulation results obtained for networks of various characteristics in terms of size, density, and intensity of broadcast data. The following modified version of the presented algorithm has been used to construct the first backbone D^0. Initially, all nodes are non-gateway nodes. Applying Step 2, Step 3 without pruning, and Step 4 of the algorithm, some node becomes an AGW. Doing so repeatedly results in D^0. In this initial phase, the priority value in a **CandidateGateway** message is not residual energy, but the number of neighbor pairs for which the neighborhood-connectivity condition does not hold (more neighbor pairs implies higher priority).

In the simulations, there have been used networks of 50, 75, and 100 nodes that are uniformly distributed over an area of 775 m × 700 m. Each node is assumed to be equipped with an omni-directional antenna. Radio propagation model follows the free-space model. For each network size, there have been considered two different levels of transmit power to

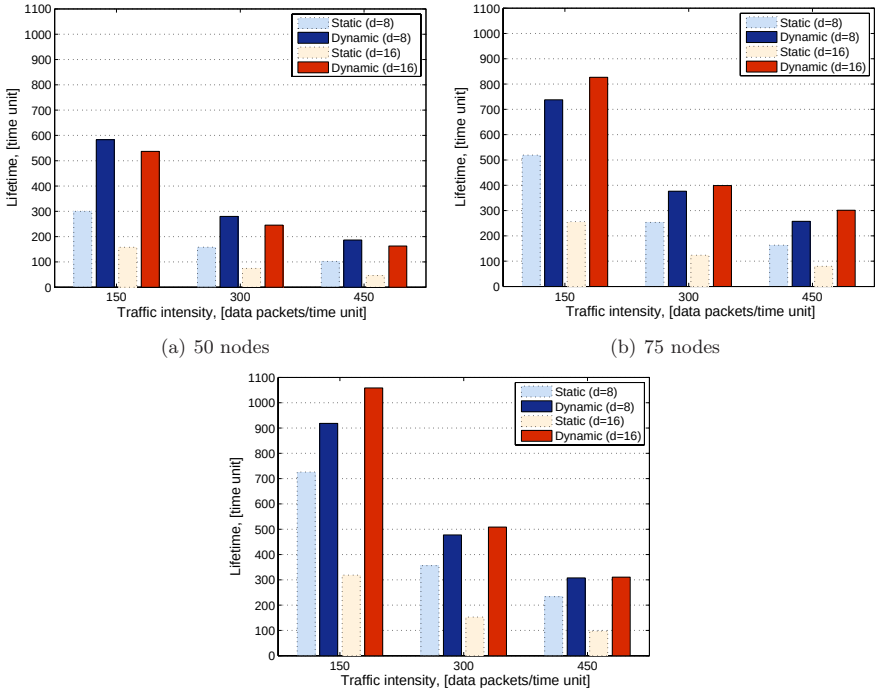
Table 10.2: Simulation results of network lifetime for distributed virtual backbone update

Number of nodes $\|V\|$	Average node degree $\bar{d}$ at t_0	Transmit power [W]	Initial backbone size $\|D^0\|$	Traffic intensity	Lifetime		
					Static	Dynamic	Ratio
50	8	0.04	17.2	150	300.0	583.2	1.94
				300	158.4	280.0	1.77
				450	102.0	186.8	1.83
	16	0.08	9.0	150	158.0	536.8	3.40
				300	74.4	245.6	3.30
				450	46.4	162.8	3.51
75	8	0.025	28.6	150	519.2	738.0	1.42
				300	253.2	376.4	1.50
				450	163.2	257.6	1.58
	16	0.05	14.2	150	256.4	826.8	3.22
				300	123.6	398.8	3.23
				450	79.6	301.2	3.78
100	8	0.018	39.0	150	726.0	918.4	1.26
				300	356.0	477.6	1.34
				450	233.6	307.6	1.32
	16	0.04	22.6	150	318.8	1058.4	3.32
				300	152.8	508.4	3.33
				450	97.6	310.8	3.18

derive network densities (i.e., average node degree) of 8 and 16, respectively. Five networks have been used for simulation for every size and density.

The focus of the simulation experiments is to examine whether and how much the network gains in terms of lifetime from the dynamic backbone. For a static backbone, energy consumption is solely due to broadcasting data messages. If the backbone is updated, some additional energy is spent for sending control messages. Because data messages are typically much longer than control messages, we assume that transmitting a data message consumes five times as much energy as transmitting a control message. There have been run simulations for three levels of traffic intensity, at which the numbers of data messages sent from a broadcasting node during one time unit are 150, 300, and 450, respectively. Table 10.2 summarizes network characteristics as well as the results of the first part of our simulation study. The results are visualized in Figures 10.7(a)-10.7(c). The first five columns in Table 10.2 display network size, average node degree, node transmit power, average size of D^0, and traffic intensity, respectively. The following three columns show, respectively, average network lifetime without and with backbone update, and the lifetime ratio. Lifetimes are given in the number of time units. The initial energy level is set to 2000 power units at every node. Moreover, $\alpha = 0.75$ and $K = 3$.

The algorithm yields lifetime gain in all scenarios in Table 10.2. In most cases the gain is significant. However, we observe that the range of improvement is quite large (from 26 % up to 378 %). If density is low ($\bar{d} = 8$), the improvement goes down when the network size increases. For example, for $\bar{d} = 8$ and a traffic intensity of 150, the improvement is close to 100 % when $|V| = 50$. When $|V|$ goes up to 100, the corresponding improvement is only 26 %. There are a couple of reasons for this. First, the amount of control traffic in relation to data traffic (which is the same in both scenarios) increases in $|V|$. Second, we have used $K = 3$ for all simulation runs. Recall that at most K nodes may be in the passive mode in one algorithm cycle, and consequently some gateways having low energy levels may not get a chance to become passive when $|V|$ is large. Our next observation is that higher node degree ($\bar{d} = 16$) leads sometimes to shorter lifetime in comparison to $\bar{d} = 8$, because the transmit



(a) 50 nodes

(b) 75 nodes

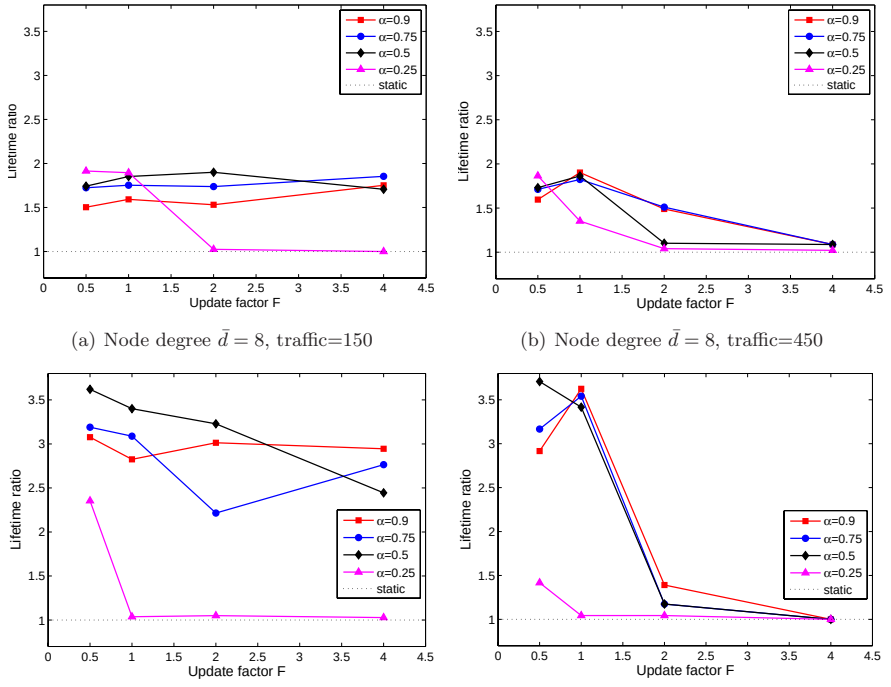
(c) 100 nodes

Figure 10.7: Network lifetime with respect to traffic intensity.

power is higher when $\bar{d} = 16$ (in order to create a dense network), while the total energy available is the same (2000 units) in both cases. The gain in lifetime, on the other hand, is significantly larger for dense networks, as in dense networks there are many alternative paths between nodes.

The second part of performance simulation aims to examine how lifetime is related to two algorithm parameters. The first is α, and the second is how often an algorithm cycle is executed. A network of 50 nodes and four scenarios with different combinations of network density (8 and 16) and traffic intensity (150 and 450 data messages per time unit) have been considered. Four values of α (0.25, 0.5, 0.75, and 0.9) have been examined. A scaling factor, denoted by F, has been used to define how often an algorithm cycle is applied. Lower F means more frequent update. If $F = 0.5$, two algorithm cycles are executed per time unit. The cases where $F = 1$, 2, and 4 correspond to one update every, every second, and every fourth time unit, respectively. The results are shown in Figures 10.8(a)-10.8(d). In the figure, the y-axis is the ratio between lifetime with backbone update and that of a static backbone.

Figures 10.8(a)-10.8(d) lead to several interesting observations. First, the threshold defined by $\alpha = 0.25$ is too low. If frequent update is allowed ($F = 0.5$), $\alpha = 0.25$ performs reasonably well, otherwise there is virtually no improvement in lifetime because the update process does not react sufficiently quickly to residual energy level. The other values of α all perform well when broadcast traffic is not intensive. We observe that, among these values, none constantly dominates the others. Second, if data traffic is intensive, α must be combined with the possibility of frequent update in order to prolong network lifetime. Third, higher network density gives, as expected, larger improvement in lifetime (except for $\alpha = 0.25$). Fourth, for intensive traffic the gain in lifetime goes down dramatically for large F. Fifth, once the update frequency is sufficiently high, increasing it further does not shorten lifetime



(a) Node degree $\bar{d} = 8$, traffic=150

(b) Node degree $\bar{d} = 8$, traffic=450

(c) Node degree $\bar{d} = 16$, traffic=150

(d) Node degree $\bar{d} = 16$, traffic=450

Figure 10.8: Lifetime with respect to α and F for a network of 50 nodes.

to any significant extent, because an algorithm cycle is executed only if there exist nodes that are or wish to become passive. All these observations suggest that frequent update should be allowed. Overall, $\alpha = 0.5$ and frequent update yield good and robust performance.

10.2 Distributed Adjustment of Transmit Power

In this section we present a distributed algorithm for maximizing network lifetime of source-independent broadcasting in stationary networks by means of transmit power control with asymmetric range assignment. In contrast to the algorithm presented in Section 10.1, here all nodes are allowed to forward broadcast messages.

10.2.1 System Model

Model Description

Unlike in Section 10.1.1, here we model an ad hoc network as a directed graph. Let the graph be denoted by $G = (V, A)$, where V and A are the sets of devices and directed links, respectively. We use p_{ij} to denote the transmit power required at node i to reach node j. Note that p_{ij} and p_{ji} are not required to be equal. Let P_i^{max} be the maximum transmit power that node i can use. The set A is defined as $\{(i,j), i,j \in V : p_{ij} \leq P_i^{max}\}$.

We assume that two nodes i and j are direct neighbors of each other, i.e., $j \in N(i)$ and $i \in N(j)$ if each of them can reach the other with a feasible transmit power level. That is, the neighborhood of node i is defined as $N(i) = \{j \in V : p_{ij} \leq P_i^{max}, p_{ji} \leq P_j^{max}\}$. Let $\hat{G} = (V, \hat{A})$ denote an undirected graph that contains one arc for each pair of direct neighbors, i.e., $\hat{A} = \{(i,j), i,j \in V : i \in N(j), i < j\}$.

The transmit power and the initial energy level of node i are denoted by P_i ($P_i \leq P_i^{max}$) and E_i^0, respectively. We use E_i to denote the residual energy at node i. The network is connected if and only if the vector of nodes' transmit power levels $P_i, i \in V$, results in (at least) one path between any pair of nodes, such that for any arc in the path, the power of the tail node is greater than or equal to the power needed to establish the arc. If message propagation is not limited by the number of hops, this is equivalent to saying that directed graph $G' = (V, A')$ defined by the power vector such that $A' = \{(i,j) \in A : p_{ij} \leq P_i, p_{ji} \leq P_j\}$ must be strongly connected. Note that we do not require symmetric power assignment. Thus the path used by i to reach j can be different from that going from j to i.

Recall that for each node, the set of its neighbors is defined for the maximum transmit power levels and is therefore static, even if nodes transmit with less power than the maximum possible. The lifetime of the network is defined as the time to the first node failure because of battery depletion. This is the time when E_i becomes zero at some node i.

Defining $N_i = \{1, \ldots, |N(i)|\}$, we let π_i, $N_i \mapsto N(i)$, be a bijection such that sequence $\left(p_{i\pi_i(1)}, \ldots, p_{i\pi_i(|N_i|)}\right)$ is monotonously non-decreasing. Hence $\pi_i(\ell)$ denotes the ℓ-th closest neighbor node i can reach. To ease the presentation, we use $p_{(i\ell)}$ as a short-hand notation for $p_{i\pi_i(\ell)}$.

Connectivity Characterization

The algorithmic idea is based on the following characterization of network connectivity.

Theorem 10.3. *Under the assumption of zero message loss and $|V| > 2$, a power vector with elements $P_i, i \in V$, gives a connected network if and only if for every node i and any ordered pair of its neighbors $j, k \in N(i)$, a message initiated at j is received by k.*

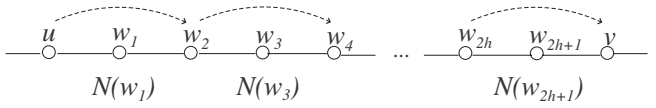
Proof. It is straightforward that the condition is necessary to guarantee that the network described by graph G' is connected. This is because a connected network implies that there exists a directed path between any two ordered pair of nodes. Next we show that the condition is also sufficient.

We define set S of nodes that have a single neighbor. Observe that the set is an independent set unless $|V| = 2$ (which falls outside the scope of the theorem). Let $\bar{S} = V \setminus S$ be the set containing the rest of the nodes, i.e., any node in $\bar{S}$ has at least two neighbors in the network. Observe that due to the neighborhood connectivity condition formulated in the theorem, any node in S has a bidirectional link in G' with a single node in $\bar{S}$.

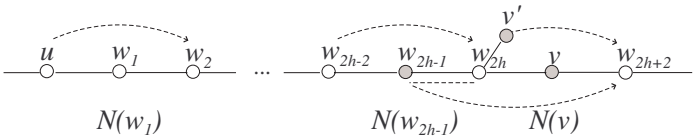
We prove by contradiction. Assume that the connectivity condition holds for any two neighbors, but the network is not connected. That is, there exist at least two nodes u and v that are not connected by a directed path in G'. Because every node in S has a bidirectional link in G' with a node in $\bar{S}$, it is sufficient to prove that there exists no such pair of nodes $u, v \in \bar{S}$ that do not receive messages of each other. Furthermore, we can limit the scope of our path search to only those involving nodes in $\bar{S}$. Note that the existence of an undirected path between any two nodes in $\hat{G}$, including u and v, is necessary for graph G' to be connected.

Consider two nodes $u, v \in \bar{S}$. Assume there exists a path of $(2h+2)$-hops from u to v in $\hat{G}$ (h is a non-negative integer). Then, due to the connectivity condition, v can receive messages from u, which is illustrated by an example in Figure 10.9(a). This contradicts our assumption that v does not receive messages from u. Consider nodes w_2 and w_4 that are neighbors of w_3. Observe that since w_2 is in $\bar{S}$, it has at least two neighbors w_1 and w_3. Node w_2 can reach w_4 by a path that involves a node from $N(w_2)$ as the first hop, and this node can be one of the following: common neighbor w_3, left neighbor w_1 of w_2, some third neighbor of w_2, or w_4 (direct connection). A dotted arrowed arc between two nodes depicts the existence of at least one such path.

Assume now that v is the end point of a $(2h+1)$-hop path in $\hat{G}$ from u to v. Node w_{2h} receives messages from u as there exists a $2h$-hop path ($h \in \{1, 2, \ldots\}$) in $\hat{G}$ from u to w_{2h}. Node w_{2h+2}, in turn, receives messages from w_{2h}. Assume that the path does not



(a) v is the end point of a $(2h+2)$-hop path from u in $\hat{G}$



(b) v is the end point of a $(2h+1)$-hop path from u in $\hat{G}$

Figure 10.9: Illustration of the neighborhood-connectivity condition for adjustable power.

involve v (otherwise v also receives messages from u). Then, node w_{2h} reaches w_{2h+2} by a path having as the first hop either w_{2h-1} (which does exist since $w_{2h} \in \bar{S}$) or some other node $v' \in N(w_{2h}) \setminus N(v)$ (which may or may not exist). Both scenarios imply that v receives messages from w_{2h} and also from u, which is illustrated in Figure 10.9(b) where grey nodes are direct neighbors of w_{2h}. Note that v' might be also a neighbor of v. In this case, however, without loss of generality, we can assume that v' coincides with w_{2h+2} (when v has exactly two neighbors), which means that v can receive messages from v' and also from w_{2h}. We have thus reached a contradiction with our assumption that v cannot receive messages from u. Hence the conclusion. □

The connectivity characterization stated in the theorem is utilized by the algorithm in order to maintain a connected network when updating transmit power for the purpose of prolonging lifetime.

10.2.2 Algorithm Description

Algorithmic Idea

Let us assume that the initial power assignment gives a connected network (for example, $P_i = p_{(i|N(i)|)}, \forall i \in V$) and power adjustment is performed at some regular time interval. However, an algorithm cycle (which implies some message overhead) is carried out only if some node is attempting to reduce its power. If such a node exists, connectivity probe messages will be used to test reachability. Let us introduce the notion of *passive node*. A node is passive when it is (tentatively) reducing its power. Assume $P_i = p_{(i\ell)}$. If node i becomes passive, it reduces the transmit power to $p_{(i\ell-1)}$ for probe messages (unless $\ell = 1$, in this case node i is not allowed to be passive). Note that even in passive mode, node i continues relaying broadcast data messages using power $P_i = p_{(i\ell)}$. As a result, the network may appear to be disconnected from the viewpoint of probe messages, but it remains connected for regular data messages.

A node becomes passive if this node considers its residual energy as low. Node i uses a reference level, denoted by $\bar{E}_i$, which is the energy level associated with the most-recent time at which i adjusted its power. Initially, $\bar{E}_i = E_i^0, \forall i \in V$. The energy level of node i, E_i, is considered as low if $E_i \leq \alpha \bar{E}_i$, where $0 < \alpha < 1$. At the beginning of a time interval, there may be multiple nodes having low energy levels. In this case an election is used to select the node with the minimum expected lifetime to become passive. The expected lifetime of node i is defined as E_i/P_i (i.e., the number of messages it can relay before battery depletion, if every message consumes P_i in energy). At most one passive node is allowed in any time interval.

Assume that node i is passive. If the network remains connected (that is, the condition in Theorem 10.3 is still true), then node i makes a definite power reduction (i.e., setting $P_i = p_{(i\ell-1)}$ for data messages as well), and becomes a normal node. Otherwise, some node will increase its power to try to "restore" connectivity. We use the term *candidate node* to refer to a node that considers increasing its power.

The operation of labeling node j as a candidate node is triggered by observing that node $j \in N(i)$ cannot reach $k \in N(i), k \neq j$. If j reaches i, but not k, then i is labeled as a candidate in addition to j, i.e., in this case both j and i are candidates. Neither of the scenarios happens if the condition in Theorem 10.3 holds at i.

Assume that the current power of a candidate node i is $P_i = p_{(i\ell)}$. Node i considers increasing its power to $p_{(i\ell+1)}$ (unless $\ell = |N_i|$, in this case i is not allowed to be a candidate node). Note that there are typically multiple candidate nodes in an algorithm cycle. Among them, one will be elected to increase its power by one step. The criterion used in node selection is the expected nodes' lifetimes. If candidate node i has $P_i = p_{(i\ell)}$, the expected lifetime, resulted from increasing power to $p_{(i\ell+1)}$, is $E_i/p_{(i\ell+1)}$. Note that a passive node may as well be a candidate node, because it may have to change the tentative power back to the original level (and becomes a normal node) to "restore" connectivity, simply due to the lack of other candidate nodes having better expected lifetime. Also, it should be pointed out that, since power is adjusted gradually in the algorithm, several algorithm cycles may be necessary before a passive node can definitely reduce its power.

An Algorithm Cycle

One cycle of the algorithm consists of several steps that are performed in a sequence. In the algorithm presentation, we assume that every node has the knowledge of its neighbors and two-hop neighbors. Figure 10.10 presents a state transition diagram for the algorithm. The entire algorithm is sketched in Algorithm III.2 and the details are outlined in Table 10.3.

Assume that Steps 1–4 are all executed in an algorithm cycle. The message overhead in each of the Steps 1, 2, and 4 corresponds to at most that of broadcasting $|V|$ data messages. In Step 3, every node does two transmissions using its maximum power. This is a rather moderate amount of overhead, and simulations show that the energy spent on control messages does pay off in terms of improved lifetime.

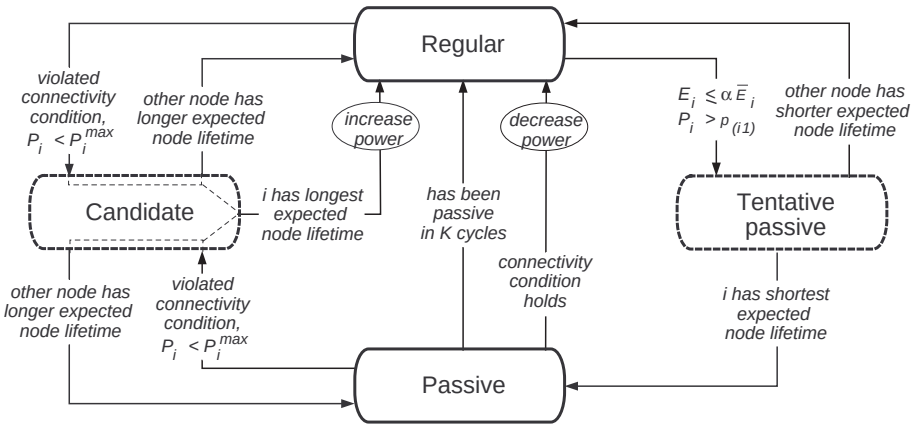


Figure 10.10: A state transition diagram for node i.

Table 10.3: The key steps of the algorithm cycle with power adjustment

Step 1 (Passive node election)
• Nodes check their residual energy levels. Node i becomes a tentative passive node, if $E_i \leq \alpha \bar{E}_i$ and $P_i > p_{(i1)}$. • Every tentative passive node broadcasts one election message containing its identification number and expected lifetime. If there is already a passive node prior to this step, this node sends an election message in which the expected lifetime is zero (in order to prevent other nodes from becoming passive). • Election messages are relayed by nodes using the same power as for regular data messages. If a tentative passive node does not see any message with lower expected lifetime, it becomes passive. In the subsequent steps, we use i^* to denote the passive node, and assume its current power, P_{i^*}, equals $p_{(i^*\ell)}$. • If no message is sent at all in this step, the algorithm cycle stops here.
Step 2 (Connectivity probing)
• Every node (including the passive one) transmits a probe message using its current power. Except for i^*, probe messages are relayed just like regular data messages. At i^*, power used to relay probe messages is $p_{(i^*\ell-1)}$. • Upon receiving a probe message, a node stores the identification number of the source. Eventually, a node has a list of sources from which messages are received.
Step 3 (Information sharing)
• The lists created in the previous step are exchanged between neighbors using maximum transmit power (i.e., node i transmits once at power P_i^{max} to distribute the list to all nodes in $N(i)$). • After the lists have arrived from the neighbors, a node transmits once more at maximum power to distribute these information. Thus there are two transmissions per node in order to enable detection violated neighborhood-connectivity condition.
Step 4 (Connectivity control and power adjustment)
• If the neighborhood-connectivity condition holds at i^*, node i^* sets $P_{i^*} = p_{(i^*\ell-1)}$, updates $\bar{E}_i$, and becomes a normal node. If a node (including i^*) detects that the condition does not hold, it marks itself as a candidate node (unless this node is already using its maximum power). • Every candidate node sends an election message. The message contains the node identification number and expected lifetime. For candidate node i having $P_i = p_{(i\ell)}$, the announced expected lifetime is $E_i/p_{(i\ell+1)}$. The election messages are relayed using the current power vector. • As a result of the election, the candidate node with the longest expected lifetime, say i', increases its power by one step and updates $\bar{E}_{i'}$. If $i' = i^*$, i^* becomes a normal node.

10.2.3 Performance Simulation

Performance simulations have been conducted for networks of different characteristics in terms of size, density, and intensity of broadcast data. As a reference solution, we consider the static scenario in which the transmit power of each node is fixed to the level needed to reach its farthest one-hop neighbor. The lifetime of this network is the lifetime of the node with the highest transmit power level. In the experiments, networks of two different sizes (25 and 50 nodes) were used with nodes uniformly distributed over an area of 400 m × 400 m. For each network size, we considered two networks with different levels of maximum transmit

Algorithm III.2 Distributed adjustment of transmit power in a stationary network

Input: V, Δ, t_0, t_{max}

$k \Leftarrow 1$
$\mathcal{P} \Leftarrow \emptyset$ // Set of passive nodes
repeat
    $t_0^k \Leftarrow t_0 + (k-1)\Delta$
    $\mathcal{C} \Leftarrow \emptyset$ // Set of candidate nodes
    $\mathcal{T} \Leftarrow \emptyset$ // Set of tentative passive nodes
    wait until $(t == t_0^k)$
    $\mathcal{T} \Leftarrow$ **Step1a**$(V \setminus \mathcal{P})$ // Select tentative passive nodes
    if $\mathcal{P} \cup \mathcal{T} \neq \emptyset$ **then**
        $i^* \Leftarrow$ **Step1b**$(\mathcal{P} \cup \mathcal{T})$ // Passive node election
        Step2(V) // Connectivity probing
        Step3(V) // Information sharing
        $\mathcal{C} \Leftarrow$**Step4a**(i^*) // Check neighborhood-connectivity condition
        if $\mathcal{C} == \emptyset$ **then**
            decreaseTXPower(i^*)
        end if
        $\mathcal{C} \overset{update}{\Leftarrow}$ **Step4b**$(V \setminus i^*)$ // Select candidate nodes
        $i' \Leftarrow$**Step4c**$(\mathcal{C})$ // Select the candidate node with the longest expected lifetime
        increaseTXPower(i')
        if $i^* \in \mathcal{C} \setminus \{i'\}$ **then**
            $\mathcal{P} \Leftarrow \mathcal{P} \cup i^*$
        end if
    end if
    $k \Leftarrow k + 1$
until $t > t_{max}$

power to model scenarios with a low and a high link density. For the sake of simplicity, we assume that each node uses an omni-directional antenna, and radio propagation follows the free-space model. However, the assumptions are not limitations of the presented algorithm.

Next we examine whether and how much we gain in network lifetime through dynamic transmit power adjustment. For the static scenario, energy consumption is solely due to broadcasting data messages. In dynamic scenarios, some additional energy is spent for sending control messages. Because data messages are typically much longer than control messages, we assume that transmitting a data message consumes five times as much energy as transmitting a control message. We performed simulations for three levels of traffic intensity, at which the numbers of data messages during one time interval are 150, 300, and 450, respectively. Table 10.4 summarizes network characteristics and the results of the simulation study. For each network size $|V|$, the table shows average node degree $\bar{d}$, node transmit power, traffic intensity, and in the following three columns show, respectively, network lifetimes for the static and dynamic power, and the lifetime ratio. Lifetimes are given in the number of time units. The initial energy level of each node is 2000 power units, and the parameter α is set to 0.8. The initial transmit power levels in the algorithm are set to the static power solution.

The algorithm yields lifetime gain in all scenarios in Table 10.4. However, for the same density, the gain is slightly smaller for the networks with 50 nodes due to higher control message overhead. To overcome this issue, the propagation of control messages in large networks can be limited in hops which is reasonable because of lower stability of long paths. We also observe that for each network size, the network with higher density has a higher lifetime gain. This can be explained by the existence of more alternative messaging routes between any two nodes, meaning that more nodes can reduce their transmit power without breaking network connectivity.

Table 10.4: Simulation results of network lifetime for distributed transmit power adjustment

Number of nodes $\|V\|$	Average node degree $\bar{d}$ at t_0	Transmit power [W]	Traffic intensity	Lifetime		
				Static	Dynamic	Ratio
25	8.16	0.025	150	537.1	1101.1	2.05
			300	268.5	532.0	1.98
			450	179.0	347.0	1.94
	15.92	0.06	150	222.5	611.0	2.75
			300	111.2	303.0	2.72
			450	74.2	167.0	2.25
50	8.72	0.013	150	1029.2	1281.0	1.24
			300	514.6	715.0	1.39
			450	343.1	498.0	1.45
	16.52	0.03	150	444.6	933.0	2.10
			300	222.3	419.0	1.88
			450	148.2	294.0	1.98

10.3 Discussion and Conclusions

We have presented two distributed algorithms for maximizing network lifetime for source-independent broadcasting in stationary networks following two different dynamic topology control approaches, namely, managing a virtual backbone and adjusting nodes' transmit power. In our simulation studies in which we have also taken into account power consumption by control messages needed to support the network topology update, we have demonstrated that both algorithms allow us to significantly prolong network lifetime. The first algorithm shows, however, better overall performance. For both algorithms, we have also examined how lifetime is affected by network density and traffic intensity.

When the first topology control approach is considered, network lifetime is prolonged by up to 251 % (compared to the static backbone) for the network with 50 nodes and up to 233 % for the network with 100 nodes in the high link density scenario. In the low link density scenario, the lifetime gain is up to 94 % and 34 % for the network with 50 and 100 nodes, respectively.

By the power-adjustment approach implemented in the second algorithm, the gain in our scenarios with the high link density is up to 175 % for the network with 25 nodes and up to 110 % for the network with 50 nodes. For the low link density scenario, the gain is smaller, up to 105 % for the network with 25 nodes and up to 45 % for the network with 50 nodes.

Comparing the results of the two topology control approaches, we observe that the first approach has a better scalability (the gain decreases slower with increasing network size as compared to the second approach). Moreover, for the same network size, the gain is larger when the first approach is applied. Another observation is that the relative lifetime gain provided by the two algorithms is insensitive to traffic intensity. Considering link density, it is expected that the achieved lifetime gain in sparse networks is smaller than that in dense networks. This is because in sparse networks the number of nodes that can support each other in terms of network connectivity is smaller.

The algorithms can be extended in several directions. Among them, the most interesting is node mobility. The second topology control approach is, however, currently less attractive for mobile ad hoc networks from the practical point of view, since dealing with non-symmetric links becomes more difficult when nodes are mobile. Considering the first algorithm as a starting point, a partial solution to dealing with mobility is to let a non-gateway node join the backbone as soon as it detects a new neighbor, but this simple approach will lead to too many AGWs. None of them attempts to prune itself before its energy level is critically low. Therefore, some additional pruning mechanism is necessary.

Chapter 11

Managing a Dynamic Virtual Backbone in Mobile Ad Hoc Networks

In this chapter, the issue of managing a dynamic virtual backbone in mobile ad hoc networks is addressed. Unlike in Chapter 10, the broadcast strategy here is not to prolong network lifetime but rather to achieve a high degree of reliability in low-power broadcast. The broadcast approach adapted for this purpose is based on managing dynamically a broadcast infrastructure represented by a virtual backbone (similar to the algorithm presented in Section 10.1). However, unlike in Section 10.1, the proposed algorithmic framework focuses on maintaining network connectivity provided by the broadcast topology. Whereas broadcast communication will unavoidably experience disturbance if the physical topology changes frequently, our framework is designed to minimize the additional impact caused by dynamic backbone management.

Beside that nodes' decisions on joining and leaving the backbone are not directly related to energy consumption, another major difference between the framework presented in this chapter and the one presented in Chapter 10 is that in the current chapter the decisions are made individually, that is, no election procedure is involved. As a result, the communication overhead is significantly reduced. Moreover, nodes do not have to synchronize their activities and therefore, can take actions using their own, local timers.

Within the framework presented in this chapter, there have been developed two distributed algorithms specifically designed to address mobility. The algorithms have the features listed in the beginning of Chapter 10. Also, the framework does not pose the restriction (e.g., [25, 42]) that the backbone must, for every two-hop-away node pair, contain one of their common neighbors. Removing this restriction allows for reducing backbone size. The presented algorithms do, however, admit control of path reliability by means of hop limit.

The major difference between the two algorithms in this chapter is that the first algorithm (presented in Section 11.1) involves probabilistic pruning, whilst the second algorithm (described in Section 11.2) is based on deterministic pruning. For both algorithms, we use the system model and the connectivity condition presented in Section 10.1.1. Note that, due to mobility, the elements of sets E and $N_k(u), k \geq 1, u \in V$ change over time which makes managing a dynamic virtual backbone an even more challenging task.

11.1 An Algorithm with Probabilistic Pruning

As has been mentioned in the beginning of the chapter, nodes' decisions on joining and leaving the backbone are made locally. At a non-backbone node, the decision is triggered by the event that connection between some of the neighbors is seemingly broken. Connectivity control is

carried out by means of probe messages. If a non-backbone node observes that some of the neighbors cannot reach each other, the node joins the backbone. Conversely, a backbone node (gateway) may consider pruning, if it detects that the neighbors are connected through some alternative backbone paths. (As will be clear later on, the node does not need to know the details of the paths.) We use a hop limit in the probe messages to allow for trading backbone size against its stability. Moreover, we show that a carefully-designed mechanism of propagating probe messages facilitates pruning.

Note that backbone connectivity is not stable if several nodes prune themselves (roughly) simultaneously. For this reason, we propose to use randomized pruning – a backbone node that is redundant in providing connectivity prunes itself with some probability (locally determined at the node). The probability increases if the node is not pruned but detects repeatedly that it is redundant. Below we detail the key elements in the algorithm.

11.1.1 Connectivity Probing

Nodes send periodically probe messages. Let Δ denote the time interval of probing, i.e., a node sends one probe message every Δ seconds. Because there is no time synchronization between nodes, every node uses its own timer to determine when the next probe message is due to be sent. A probe message has the format

$$\textbf{Probe}(Source, FirstGW, TimeStamp, TTL),$$

where *Source* identifies the source node, $FirstGW$ is the first gateway node that has relayed the message, $TimeStamp$ is a time stamp filled by the source (to make it possible to distinguish between **Probe** messages originated from the same node), and TTL is the time-to-live field. When a message is transmitted by the source, the field $FirstGW$ is empty. Every node receiving a **Probe** message uses the information in the message for updating its local database. Non-backbone nodes never retransmit **Probe** messages, whilst gateway nodes, depending on the information in their local databases, either retransmit the **Probe** message (with updated TTL and in some cases $FirstGW$ fields) or discard it.

The TTL field controls the scope within which **Probe** messages are propagated. A large initial TTL (in the number of hops) corresponds to allowing the use of long paths to connect nodes. Long paths are, however, not stable in the presence of node mobility. Therefore a relatively small hop limit (a few, but a minimum of two, hops) should be used. In our algorithm, setting the initial TTL to k, in effect, means that the neighbors of a node must be connected by backbone paths of no longer than k hops. If such paths cannot be established, then the node should join the backbone to create two-hop paths for its neighbors.

Upon receiving a **Probe** message, a node uses the information in the message to update a database. (We discuss further details of this database in Section 11.1.2). If the node (denoted by u) is a gateway node, it takes the following additional actions.

- If u is a direct neighbor of *Source*, u checks whether the field $FirstGW$ is empty. If this is the case (i.e., the message comes directly from the source), u reduces the value of TTL by one and transmits the message $\textbf{Probe}(Source, u, TimeStamp, TTL)$, otherwise u discards the message.
- If u is not a direct neighbor of *Source*, u does the following.
  - If u has previously received a **Probe** message containing the same *Source*, $FirstGW$, and $TimeStamp$, but a TTL higher than or equal to that in the current message[1], u discards the message.
  - Otherwise u sets $TTL = TTL - 1$. If $TTL > 0$, u transmits the message $\textbf{Probe}(Source, FirstGW, TimeStamp, TTL)$.

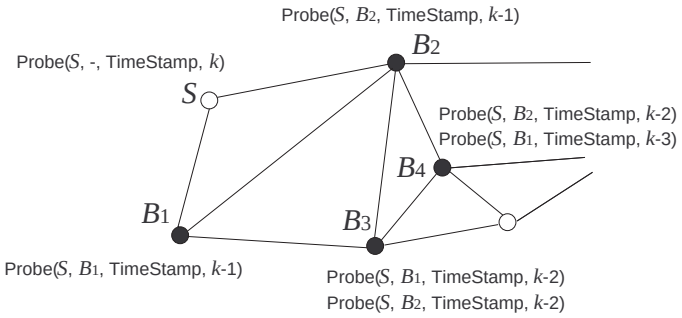


Figure 11.1: An illustration of connectivity probing.

The behavior of gateway nodes in relaying **Probe** messages is illustrated in Figure 11.1, where gateway nodes are marked black. The hop limit k is assumed to be greater than three. In this example, node S initiates a **Probe** message. This message, with some of the fields being updated, is retransmitted once by each of the one-hop gateways B_1 and B_2. The message relayed by B_2 is further relayed by B_4. The same message reaches B_4 also via B_3, but is not relayed as the TTL value is smaller. On the other hand, node B_4 does relay the message that passed through B_1 and then B_3. Now consider node B_3. It relays the message from B_1 and that from B_2, but not that arriving via B_4.

Every node sends one **Probe** message during a period of Δ seconds. Let $B(u)$ denote the set of backbone neighbors of u, and $B_k(u)$ the set of gateway nodes within the k-hop environment of u. A **Probe** message initiated at u is retransmitted by every node in $B(u)$. Each of these retransmitted messages will be relayed by at most $|B_k(u)|-|B(u)|$ nodes, if delay increases with the number of hops. Therefore, the total number of transmissions of **Probe** messages (in a static network) does not exceed $\sum_{u\in V}(|B(u)| + |B(u)| \cdot (|B_k(u)| - |B(u)|))$ within one time period. In a mobile network, there can be some changes in the backbone during the time interval Δ since nodes' decisions of joining and leaving the backbone are not synchronized. Therefore, the estimated message overhead can be viewed as an average expected amount of overhead for a mobile network.

Detecting the Existence of Alternative Paths between Two-hop Neighbors

If we were only interested in verifying whether or not neighborhood connectivity holds, it is sufficient to let a gateway node relay the **Probe** message of a source *once*. The objective of the design of the connectivity probing procedure that has been described in this section is to ensure correct pruning. As a result of the probing scheme, a gateway node must realize that it is redundant *if and only if* its neighbors are connected through other gateway nodes forming some path of at most k hops. The correctness of the probing scheme is illustrated by the following examples.

Consider a backbone B that connects its one-hop neighbors u and v that do not have a direct link with each other (see Figure 11.2(a)). Assume there exist an alternative path connecting u and v and contains B_1 and B_2 (and possibly some additional gateway nodes between them). We show that a gateway node in the path, which is not the first gateway to u, must relay a **Probe** message of u multiple times if u has more than one gateway in its one-hops neighborhood. Suppose that gateway B_2 receives a **Probe** message from u via B and via B_1. Then, if all gateway nodes relay the message only once, and B_2 first receives the message from u relayed by B (quite likely due to a smaller number of hops), then node v will

[1]Node u consults its database (Section 11.1.2) to perform this check.

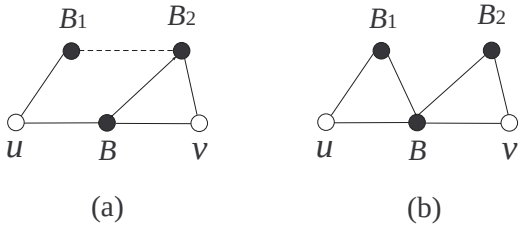


Figure 11.2: Detecting the existence of alternative path.

be unaware of the fact that it is connected to u via some path not going through B. For this reason, B_2 should relay the same **Probe** message, if it comes from a different first-hop node (B_1 in the example).

Next we demonstrate that a gateway node directly connected to *Source* should relay the **Probe** message no more than once. Assume node u is connected to v through two paths, both involving B (see Figure 11.2(b)). If B forwards the **Probe** message of u received through the direct link and also the one relayed by B_1, v will erroneously believe it is reached by u via a path not passing through B, and B is therefore eligible for pruning. Storing the entire path in a **Probe** message would have resolved this scenario. However, the solution of letting B forward the **Probe** message only if it comes directly from u is more efficient as the message size can be kept constant.

11.1.2 Maintaining a Connectivity Information Database

A node uses a database to store connectivity information. An entry in the database contains the fields *Source*, $FirstGW$, $TimeStamp$, and TTL. The fields have the same meaning as those in a **Probe** message. At most one entry can be present for each combination of *Source* and $FirstGW$. In addition, an entry is associated with a timer, denoted by $Time$. This is the time elapsed since a **Probe** message corresponding to the entry has been received. When a node receives a **Probe** message, the node checks whether both *Source* and $FirstGW$ are present in the database. If not, a new entry is created (with $Time = 0$). If *Source* and $FirstGW$ are present but the associated $TimeStamp$ is older than the one in the message, the $TimeStamp$ is updated and $Time$ is reset to zero. In case *Source*, $FirstGW$, and $TimeStamp$ in the message are identical to those of an entry, but the message contains a higher TTL, then the TTL field of the entry is updated (but $Time$ is not reset).

When $Time$ exceeds some threshold, the entry becomes invalid, that is, the connection from *Source* via any path using $FirstGW$ as the first hop is considered lost. An invalid entry is removed. Ideally, a **Probe** message corresponding to an entry should arrive at most every Δ seconds under a static topology. However, because delay is variable, the threshold should be set to $\alpha_1\Delta$ for some $\alpha_1 > 1$.

In addition to information collected via **Probe** messages, a database contains entries for directly connected one-hop neighbors. These entries are not updated by **Probe** messages. Once a (new) neighbor is detected, an entry for the neighbor is added. Both *Source* and $FirstGW$ are set to the identification of this neighbor. The entry is kept alive as long as the neighbor relation lasts.

A quite small k makes information storage scalable, as the database of a node will, in the worst case, contain source nodes within a local k-hop environment. More specifically, the number of entries at node u, at its maximum, equals $|N(u)|+\sum_{v\in N_k(u)\setminus N(u)}|B(v)|$. Moreover, it can be easily realized that, in order to control neighborhood connectivity, a node needs to store entries for source nodes that are at most two hops away in the underlying graph. Therefore the database size is very small under the assumption that nodes have knowledge

of their two-hop neighbors. This is, in fact, a rather standard assumption in ad hoc networking. However, it should be remarked that collecting this information in practice introduces some additional overhead, especially in highly mobile networks. We end our discussion of connectivity information database by formalizing the correctness of the algorithm in terms of controlling neighborhood connectivity.

Theorem 11.1. *Consider node u and any pair of its neighbors v and w that have no direct link with each other. Under static network topology and zero message loss, the scheme of connectivity probing leads to the following results.*

1. *There will be an entry where $Source = v$ and $FirstGW = u$ in the database of w, and vice versa, if and only if u is a gateway node.*
2. *When u is a gateway, there will be an entry where $Source = v$ and $FirstGW \neq u$ in the database of w, and vice versa, if and only if there exists an alternative backbone path (i.e., a path not passing through u) connecting v and w using at most k hops.*

Proof. That the first statement is valid is obvious from the fact that only gateways retransmit **Probe** messages (see Section 11.1.1). The necessity of the if-condition in the second statement follows from the fact that if the database of w has an entry where $Source = v$ and $FirstGW \neq u$, then there must be a path from v to w, and the path should not involve gateway u (otherwise, the **Probe** message received by u not directly from v would be discarded by u). Next, we show that the condition is also sufficient. Assume there exists a path between v and w which has u' as the first gateway and the path does not pass u. Then the **Probe** message from v will be retransmitted by u'. If w is a direct neighbor of u', it receives the message directly from u' since u' always retransmits messages directly coming from v. Otherwise, the path contains other gateways that connect v and w, and each of these gateways discards the received **Probe** message only if it has already retransmitted the same **Probe** (i.e., with the same $Source$, $FirstGW$ and $TimeStamp$ fields) with a TTL higher than or equal to that in the received **Probe** message. Therefore, the **Probe** message sent by v will be finally received by w which will accordingly update its database. Hence the conclusion. □

11.1.3 Information Exchange and State Update

Connectivity information databases are shared between neighbors. Due to the multicast advantage, a node can deliver the same information to its neighbors in one transmission, meaning that the amount of overhead imposed by exchanging databases is small. Sending the database to neighbors is triggered by two events. The first is when the time elapsed since the previous information exchange reaches a threshold $\alpha_2\Delta$, where $\alpha_2 \geq 1$. The second event is the detection of new neighbor(s) – whenever a new neighbor is detected, the database is sent regardless of the time since the previous exchange.

A non-backbone node makes a decision of joining the backbone if there is seemingly no connection between some of its neighbors. More specifically, when a non-backbone node u receives the database of v, u examines whether each of the other neighbors appears as $Source$ in at least one entry. If this is true, u remains a non-backbone node. Otherwise u changes its state and becomes a gateway node.

Pruning is considered at a gateway node for which alternative paths exist between its neighbors. Typically, there are multiple backbone nodes that are mutually dependent, that is, a gateway node is redundant only if some other nodes (each of which is also considered redundant) remain in the backbone. Simultaneous pruning of several nodes may lead to disconnectivity and thus unreliable broadcast. We propose to deal with this issue by randomized pruning. Let $\beta\Delta$ ($\beta > 0$) denote the frequency of executing pruning, i.e., every $\beta\Delta$ seconds a gateway node examines whether it is eligible for pruning, and if so, applies pruning probabilistically. The eligibility of being pruned is derived from the databases that have

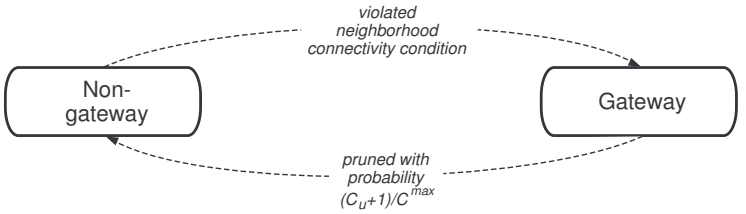


Figure 11.3: A state transition diagram for managing a virtual backbone in mobile ad hoc networks with probabilistic pruning.

been obtained most recently from the neighbors. A gateway node u is eligible for pruning, if for all $v, w \in N(u)$, the database of v contains (at least) one entry where $Source = w$ and $FirstGW \neq u$, (i.e., v and w are connected through a path not involving u).

An eligible gateway node u prunes itself with probability $(C_u + 1)/C^{max}$, where C_u is a counter that keeps track on the number of consecutive pruning attempts. Every time u detects it is not eligible, it sets $C_u = 0$. If node u is not pruned but remains eligible after $\beta\Delta$ seconds, it sets $C_u = C_u + 1$. The maximum number of consecutive unsuccessful attempts, i.e., before u will definitely prune itself, is thus $C^{max} - 1$.

Note that there is no election or synchronization of any kind in the process of making decisions of joining and leaving the backbone. For a non-backbone node, a decision of changing the state is solely based on information that are regularly obtained from neighbors. A gateway node combines information coming from neighbors with probabilistic pruning. Thus state update does not involve any extra message overhead. Figure 11.3 shows a state transition diagram of the described algorithm.

11.1.4 Performance Simulation

There have been conducted performance simulations for an ad hoc network operating in the 2.4 GHz band and consisting of 50 mobile nodes. Initially, all nodes are non-backbone nodes and are uniformly distributed over an area of 750 m × 700 m. Nodes' mobility is modeled by the random waypoint model [24] with constant speed ν (the same for all nodes) and the waiting time of 1 sec.

It is assumed that each node uses an omni-directional antenna with 0 dBi antenna gain, and distance-dependent radio propagation follows the free-space model (Friis transmission equation with the system loss factor of one and receivers' antenna gain being set to zero). The receive signal threshold, which defines whether there exists a direct link between two nodes or not, is set to 10^{-7} mW (−70 dBm).

In the beginning all nodes join the network simultaneously. This is not a limitation of the proposed algorithm which is designed to deal with non-synchronized behavior of nodes. Note, however, that due to an immediate database exchange after a new link is established, a node that newly joins the network will most probably cause one or few of its neighbors to go into the gateway state. Thus if many nodes join the network in a very short time period, the backbone size will increase significantly, after which the backbone will smoothly adapt itself to the (new) topology by self-pruning.

Two network density scenarios are considered. In the low-density scenario, all nodes transmit at 30 mW, and the average node degree of the initial topology is 5.68. In the high-density scenario, the transmit power of each node is 60 mW, and the initial average node degree is 10.72. For each network density, we experiment with two different speed levels (3 m/s and 10 m/s). By simulating the network behavior during 300 sec with the smallest simulation time unit of $\delta = 0.2$ sec. The setting of the algorithm parameters used in simulations is shown

Table 11.1: Parameters of the algorithm with probabilistic pruning

Parameter	Notation	Value
Connectivity probing interval	Δ	1 sec
Maximum age factor of a database entry	α_1	1.4
Frequency factor of connectivity information exchange	α_2	1
Initial TTL of a **Probe** message	k	4
Pruning frequency factor	β	1
Maximum pruning attempts	C^{max}	7

in Table 11.1. As we will show later, the parameters can be adjusted to network density and mobility to achieve better performance. However, in this part of the study we use the same setting in all test scenarios.

No transmission delay is assumed at a source node and a random transmission delay is assumed at a relaying node. A node may delay relaying a message for δ seconds with 50 % probability. The link delay is assumed to be constant per transmission and is equal to δ. Thus, the minimum and the maximum total delay of a message received at hop k (until the message will eventually be sent by a relaying node) is δk and $\delta(2k-1)$, respectively.

To measure the algorithm performance, we use the backbone size and two ratios. The first ratio, denoted by r_1, is the proportion of node pairs that are connected by the backbone in relation to the total number of node pairs computed as $\frac{1}{2}|V|(|V|-1)$. The second ratio is denoted by r_2 and shows the degree of connectivity of the underlying graph G, that is, the number of connected node pairs in the graph in relation to the total number of node pairs. This ratio represents the maximum achievable connectivity degree. Occasionally, r_2 drops below 1.0 when graph G becomes disconnected due to node mobility. Observe that $r_1 \leq r_2 \leq 1$ always holds. Thus, the ability of the algorithm in dealing with node mobility is illustrated by how well the curve of r_1 follows that of r_2.

Figures 11.4 and 11.6 present the simulation results for $\nu = 3\,\mathrm{m/s}$ in the low- and the high-density scenario, respectively. Figures 11.5 and 11.7 show the simulation results for $\nu = 10\,\mathrm{m/s}$. We observe that the difference between the two ratios does not exceed 5 % in the high-density scenarios and only at a few time points exceeds 10 % in the low-density scenarios. Moreover, the curves of the two ratios coincide during most of the simulation time in all four cases. The smallest (in average over time) backbone size is achieved in the high-density scenario when $\nu = 3\,\mathrm{m/s}$ (see Figure 11.6), whereas the largest backbone can be observed in the low-density scenario when $\nu = 10\,\mathrm{m/s}$ (see Figure 11.5).

Figure 11.8 demonstrates the network dynamics over time in the low-density scenario when $\nu = 10\,\mathrm{m/sec}$. The figure is a zoomed-in version of Figure 11.5(a). We have chosen the time interval $[165\,\mathrm{sec}, 185\,\mathrm{sec}]$ and marked four points. At point 1, the graph becomes disconnected, but all pairs of nodes that can be connected in the graph are connected through the current backbone, i.e., $r_1 = r_2$. In the next second (point 2), ratio r_1 slightly drops due to mobility in the larger subnetwork in Figure 11.8(a). The network connectivity is at its maximum possible level at point 3 (see also Figure 11.8(a) where dashed lines are the radio links, and the backbone nodes are marked by squares). In the next second, the graph becomes connected again, and the backbone tries to adjust itself to the new topology. Figure 11.8(b) displays the network topology in second 177 (point 4). At this time point, the network and the underlying graph become fully connected again.

In the next part of our study we investigate the effect of parameter setting on the algorithm performance. The low-density scenario and $\nu = 3\,\mathrm{m/sec}$ have been chosen for the experiment. Table 11.2 shows the average backbone size for different combinations of the maximum age factor α_1 and the maximum number of pruning attempts C^{max}. Due to probabilistic pruning, the backbone can be different at the same time point in different simulation runs, even if the network and the parameter setting remain the same. Therefore, for each parameter setting, we present the average result after five simulation runs. In each simulation run, we compute the

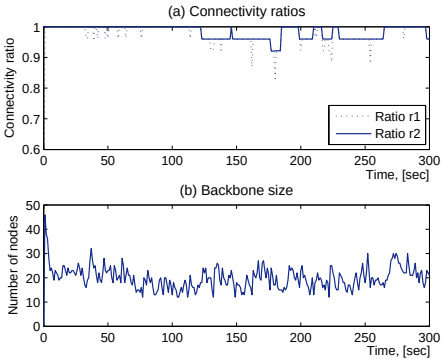


Figure 11.4: Dynamic backbone statistics (low density, $\nu = 3\,\mathrm{m/sec}$).

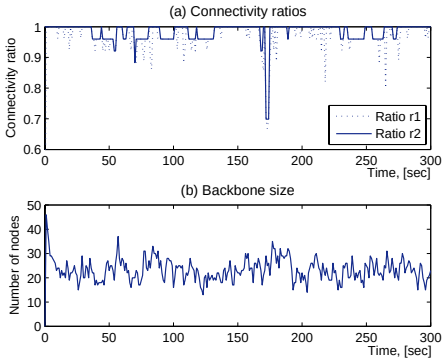


Figure 11.5: Dynamic backbone statistics (low density, $\nu = 10\,\mathrm{m/sec}$).

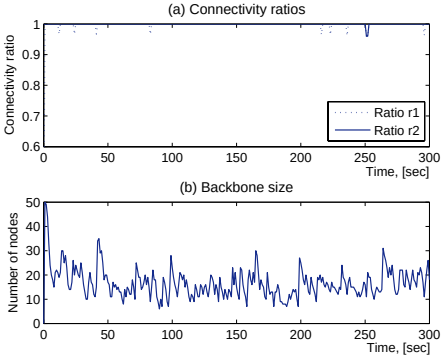


Figure 11.6: Dynamic backbone statistics (high density, $\nu = 3\,\mathrm{m/sec}$).

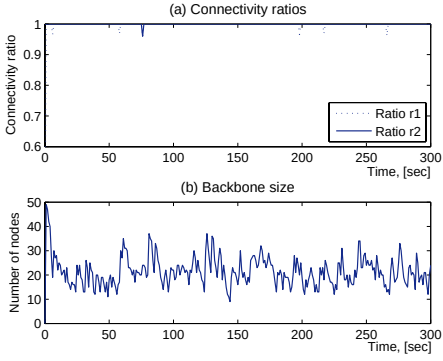


Figure 11.7: Dynamic backbone statistics (high density, $\nu = 10\,\mathrm{m/sec}$).

average backbone size over time starting from the moment when full network connectivity (by the backbone) has been achieved for the first time, i.e., we exclude the initial phase (typically one or two seconds). Table 11.3 presents the average percentage (after five runs) of the simulation time when $r_1 = r_2$, i.e., when the network connectivity through the backbone is the maximum achievable connectivity with respect to the underlying graph.

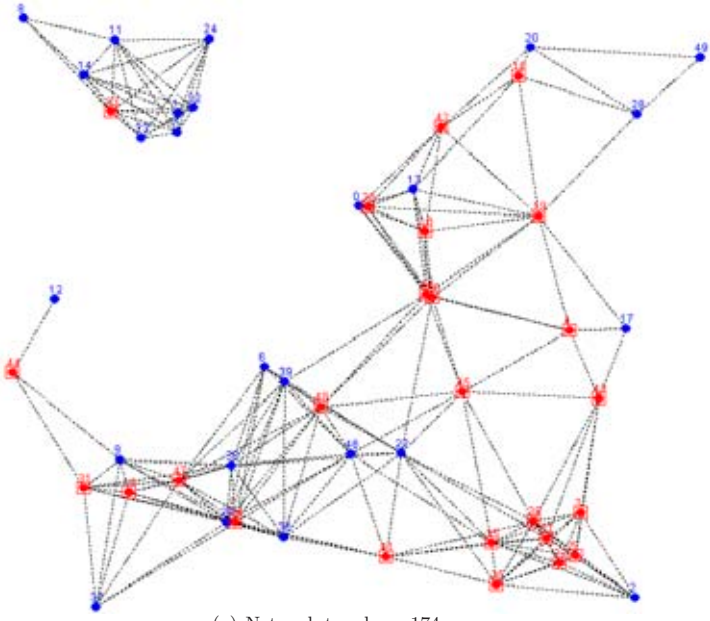
We observe that more conservative pruning makes the broadcast infrastructure more reliable at the cost of a larger backbone. This is due to a smaller probability of simultaneous pruning of backbone nodes. A lower value of α_1 allows the network to faster adjust to network topology changes. However, this also increases the size of the backbone. Moreover, the effect of the maximum age factor on the backbone size is greater than that of the maximum number of pruning attempts.

Table 11.2: Average backbone size

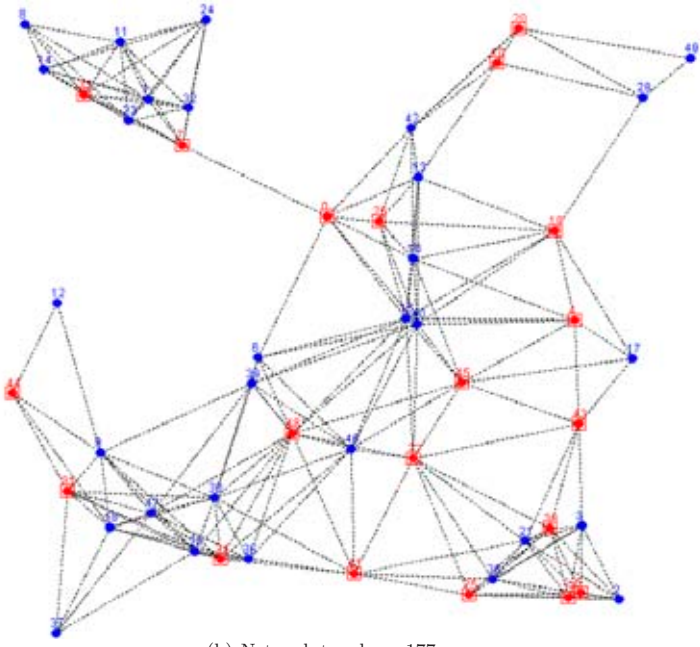
α_1	C^{max}		
	3	7	11
1.2	25.86	27.83	29.84
1.4	19.28	20.11	21.20
1.6	17.41	18.59	18.81
1.8	17.54	18.07	18.40

Table 11.3: Full connectivity duration

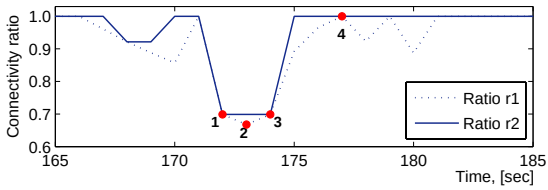
α_1	C^{max}		
	3	7	11
1.2	94.87	96.73	97.60
1.4	87.27	91.53	92.40
1.6	82.73	86.93	89.27
1.8	73.53	80.80	84.67



(a) Network topology, 174 sec.



(b) Network topology, 177 sec.



(c) Dynamics of the connectivity ratios

Figure 11.8: Dynamic behavior of the algorithm (low density, $\nu = 10\,\mathrm{m/sec}$).

11.2 An Algorithm with Deterministic Pruning

The algorithm presented in this section improves the algorithm described in Section 11.1. One improvement is the use of deterministic pruning that is more reliable than randomized, probabilistic pruning. Another improvement is better message efficiency. In the algorithm based on probabilistic pruning (Section 11.1), a probe message of a node may lead to several messages, each of which has to be relayed by the backbone. This inefficiency is resolved in the algorithm presented in this section. The aforementioned improvements, however, may come at a cost of a larger backbone size, which is studied in Section 11.2.4.

Similar to the algorithm presented in Section 11.1, here we also distinguish between non-backbone nodes and gateway nodes. In addition, we assume that a gateway node can be in two states: a regular gateway (RGW) and a pruning gateway (PrGW). In the pruning state, a gateway attempts to prune itself from the backbone. There may be (and in fact, should be) multiple PrGWs. Note that both RGWs and PrGWs relay broadcast data, that is, for data traffic there is no difference between RGW and PrGW. In discovering reachability, however, RGWs and PrGWs behave differently in relaying **Probe** messages, as explained in the next section.

11.2.1 Connectivity Probing

Connectivity probing is performed by means of probe messages. Each node broadcasts probe messages regularly with time interval Δ (using a local timer) via the current backbone. A probe message has the format

$$\textbf{Probe}(Source, TimeStamp, TTL, Attribute),$$

where *Source* is the source node of the broadcast message, $TimeStamp$ is a time stamp filled by the source, TTL is the time-to-live field, and *Attribute* is a binary field which is *true* if the message has been relayed by at least one PrGW and $false$ otherwise. The initial value of the *Attribute* field is always $false$. Before relaying a **Probe** message, a PrGW sets the value of this field to *true*. Note that *Source* and $TimeStamp$ together give a unique identification to the message. Similar to the algorithm presented in Section 11.1, the TTL field is initially set to k, where k is a configurable parameter, and is used to control the scope of a **Probe** message.

Both RGW and PrGW relay **Probe** messages, but their treatments of the *Attribute* field differ. An RGW/PrGW uses a buffer to store recently received **Probe** messages. For each message in the buffer, the node stores the time at which the message was most recently relayed. Moreover, an RGW/PrGW records the time (if any) when it changed its state to RGW/PrGW. When a **Probe** message arrives, a decision on whether or not to relay the message is made by the RGW/PrGW according to the sequence described in Table 11.4.

As can be seen from Step 1, similar to the algorithm presented in Section 11.1, direct neighbors to the source node in this algorithm also retransmit the **Probe** messages only once. In Step 3, t_s and t_m denote the time of the most recent state change and the time recorded in the buffer, respectively. As the result of this step, the node considers relaying a message in case the message was relayed before but after that the node changed its state.

11.2.2 Maintaining a Connectivity Information Database

Every node has a database containing sources (within k hops) from which **Probe** messages have been received. A node keeps at most two entries (of different attributes) for any source. A database entry is associated with a timer. An entry is removed if the timer gets expired. An entry is specified by *Source*, $TimeStamp$, *Attribute* and TTL. When an entry is timed out, connection (of type specified by *Attribute*) with *Source* is considered lost. Upon receiving a **Probe** message, a node updates its database according to the steps described in Table 11.5.

Table 11.4: Connectivity probing

Probe message rebroadcast decisions at RGWs and PrGWs	
Step 1 (*Neighbor check)*	If the node that appears as *Source* is a direct neighbor to the receiving, discard the message if it does not arrive from the source (i.e., has $TTL < k$) or go to Step 5 otherwise.
Step 2 (*Buffer check)*	If the buffer does not contain a message having the same *Source*, *TimeStamp*, and *Attribute*, go to Step 5.
Step 3 (*Time check)*	If $t_s > t_m$, go to Step 5.
Step 4 (*TTL check)*	If the arriving message has a smaller TTL, discard the message.
Step 5 (*TTL update)*	Set $TTL = TTL - 1$. If TTL is zero, discard the message.
Step 6 (*Relay)*	If the node is an RGW, retransmit the message, otherwise (i.e., the node is a PrGW) retransmit the message with *Attribute* set to *true*. Update the buffer.

Table 11.5: Connectivity information database update

Processing of a received **Probe** message	
Step 1	If *Source* refers to the node itself, discard the message.
Step 2	If the *Source* and *Attribute* combination in the message is not present in any entry of the database, a new entry is created, and the values of the fields in the message are copied. A timer is then set for this entry.
Step 3	If the database contains an entry having the same *Source* and *Attribute*, the node examines if the message contains a newer time stamp. If so, the *TimeStamp* field is updated, and the timer is reset.
Step 4	In case *Source*, *Attribute*, and *TimeStamp* in the message are identical to those of an entry, but the message contains a higher TTL, the TTL field of the entry is updated (but the timer is not reset). This scenario occurs if the delay of a shorter path happens to be greater than that of a longer one.

Again, similar to the algorithm presented in Section 11.1, a database contains entries for directly connected one-hop neighbors, and these entries are not updated by **Probe** messages but once a (new) neighbor is detected. The entry is kept alive as long as the neighbor relation lasts.

Let $\mathcal{P}_k(v, u)$ denote the set of backbone paths connecting nodes v and u with at most k hops. A path in $\mathcal{P}_k(v, u)$ is called an RGW path if it consists of RGWs only, and a PrGW path if it contains at least one PrGW. Theorem 11.2 formalizes the following properties of nodes' databases with respect to non-neighbor nodes.

Theorem 11.2. *Under the assumptions of zero message loss and no jitter, the proposed schemes of reachability probing and database update have the following properties in a stationary network with a connected backbone.*

1. *For Source $= v$, the database of $u \notin N(v)$ will contain a single entry with Attribute $=$ false, if and only if all paths in $\mathcal{P}_k(v, u)$ are RGW paths.*
2. *For Source $= v$, the database of $u \notin N(v)$ will contain two entries with Attribute $=$ false and Attribute $=$ true respectively, if and only if $\mathcal{P}_k(v, u)$ contains both RGW and PrGW paths.*

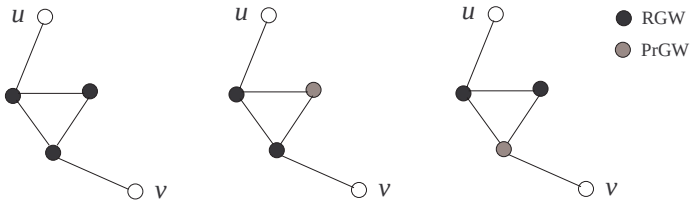


Figure 11.9: Scenarios of backbone connectivity between two nodes.

3. *For Source $= v$, the database of $u \notin N(v)$ will contain a single entry with Attribute $=$ true, if and only if all paths in $\mathcal{P}_k(v,u)$ are PrGW paths.*

Proof. Assume the database of u contains a single entry with $Attribute = false$. This implies that the **Probe** message of v was relayed by RGWs only in all paths in $\mathcal{P}_k(v,u)$ (otherwise, the *Attribute* field would be changed, and an entry with $Attribute = true$ for the same source would be added). Now assume that all paths in $\mathcal{P}_k(v,u)$ are RGWs. This means that node u never receives **Probe** messages of v with $Attribute = true$ and therefore, its database cannot contain an entry where $Source = v$ and $Attribute = true$, but there will be an entry where $Source = v$ and $Attribute = false$. Furthermore, there can be only one such entry since if the database already contains an entry for the same *Source* and the same *Attribute*, it can only be either removed (which will not happen under the assumptions of zero message loss and no jitter) or updated. If such an entry does not exist for $Source = v$, it will be created by u once it receives a **Probe** message of v. We have thus shown that the first item is valid.

Assume now that $\mathcal{P}_k(v,u)$ contains PrGW paths only. This means that when the **Probe** messages of v follows any of the paths in $\mathcal{P}_k(v,u)$, its *Attribute* field will be always changed since all paths contain at least one PrGW. As a result, all **Probe** messages of v received by u will always contain $Attribute = true$, which means that the database of v cannot contain entries where $Source = v$ and $Attribute = false$ and will contain a single (see the proof of the first item) entry where $Source = v$ and $Attribute = true$. The necessity of the condition that all paths in $\mathcal{P}_k(v,u)$ are PrGW paths in the third item of the theorem is straightforward, which completes the proof of the third item.

The proof of the second item of the theorem follows from the proofs of the first and the third items and is straightforward. Hence the result of the theorem. □

Since links are bi-directional, in each of the scenarios listed in Theorem 11.2 the database of v will also contain the corresponding entry/entries for source u. The three scenarios of the theorem are illustrated in Figure 11.9.

11.2.3 Information Exchange and State Update

Like in the algorithm presented in Section 11.1, every node shares its connectivity information database with one-hop neighbors utilizing the wireless multicast advantage. A node delivers the database to its neighbors at regular time intervals using its local timer. In addition, every time a new link is established (i.e., two nodes become neighbors of each other), there is an immediate database exchange between the two nodes.

There are four types of state transition: from non-gateway to RGW, from PrGW to non-gateway, from PrGW to RGW, and finally from RGW to PrGW. The first three types take place deterministically (i.e., using pre-defined conditions), whereas going from RGW to PrGW occurs in a randomized manner. The state transition diagram is demonstrated in Figure 11.10.

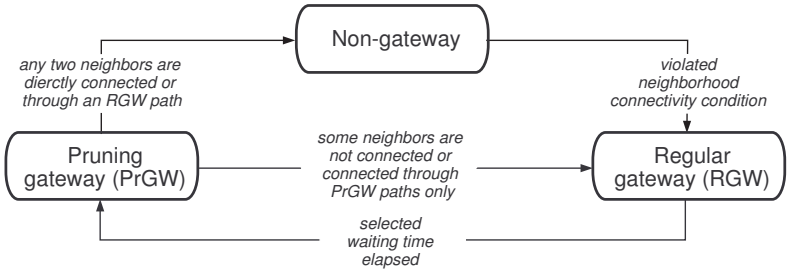


Figure 11.10: A state transition diagram for managing a virtual backbone in mobile ad hoc networks with deterministic pruning.

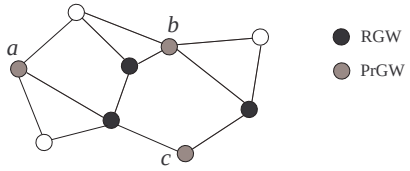


Figure 11.11: An example of redundant PrGWs.

A non-gateway node joins the backbone if some of its neighbors do not reach each other. Thus, each time the node receives the database from a neighbor, it examines if all the other neighbors are present as *Source* (regardless of the *Attribute* field) in the database. If not, the non-gateway node joins the backbone and changes its state to RGW.

A PrGW is said to be redundant if pruning this node preserves the CDS property of the backbone. However, simultaneous (or roughly simultaneous) pruning of several PrGWs, each of which is redundant, may break backbone connectivity. In Figure 11.11, three PrGWs, a, b, and c, are all redundant. We can prune a and b or a and c, whereas the backbone is no longer a CDS if pruning takes place at both b and c.

A PrGW prunes itself if its neighbors are either directly connected, or connected via RGW paths. It is easy to realize that pruning under this condition ensures the CDS property of the backbone. Referring again to Figure 11.11, any two neighbors of a are connected either directly or via some RGW path. Thus a is allowed to prune itself. On the other hand, for b and c, there exists some neighbor pair connected through PrGW paths (via b or c) only, meaning that none of the two nodes will prune itself. Note that it is in fact safe to prune one of the two nodes. However, because nodes do not coordinate their actions, a pruning decision at any of the two nodes implies a risk of breaking connectivity.

Assume that a node became PrGW at time t_s. The node applies the procedure described in Table 11.6 to determine its next state. For the example in Figure 11.11, PrGW a will prune itself, whereas both b and c will become RGWs. Because a randomized approach is used for state transition from RGW to PrGW, it is likely that later on b and c will choose different time moments to enter the PrGW state again. In this case, another pruning will take place.

An RGW enters the pruning state (i.e., becomes PrGW) from time to time in order to examine whether or not it can prune itself from the backbone. Each RGW decides by itself when state transition to PrGW will take place. To keep backbone size small, state transition from RGW to PrGW should be frequent. On the other hand, if RGWs change their state to PrGW too aggressively, they will most likely block each other's pruning operation. (This

Table 11.6: Possible state transitions for a PrGW node

Step 1	Each time it receives the database from a neighbor, it examines the entries containing the other neighbors as *Source*. For every entry having $Attribute = false$ and a time stamp later than t_s, the corresponding pair of neighbors is marked. Also, all pairs of nodes that are direct neighbors to each other are marked. Once all pairs become marked (which requires in most cases that more than one database have been received), the PrGW prunes itself and becomes a non-gateway node.
Step 2	If for any neighbor pair, the database of one of them contains only one entry of the other, where the time stamp is later than t_s and $Attribute = true$, the PrGW changes its state to RGW.
Step 3	If the database of a neighbor does not contain any entry of another neighbor, the PrGW changes its state to RGW. Note that this scenario may occur only if at least one of the neighbors is new to the PrGW, i.e., a link has just been established.

happens to PrGWs b and c in Figure 11.11.)

We combine randomized time selection and a back-off mechanism for state transition from RGW to PrGW. An RGW uses a parameter t which is restricted to lie between t_{min} and t_{max}. The RGW waits for an amount of time that is uniformly distributed in $[t_{min}, t]$, before it changes its state to PrGW. When a non-gateway node becomes RGW, it sets $t = \alpha \cdot t_{min}$. Here α ($\alpha > 1$) is a pre-defined parameter. After each unsuccessful attempt of pruning (i.e., the node changes its state from PrGW back to RGW), the node performs a back-off by setting $t = \min\{\alpha \cdot t, t_{max}\}$. As a result, if two (or more) PrGWs block each other's opportunity of pruning, the probability that they again become PrGWs simultaneously will decrease.

11.2.4 Performance Simulation

The results of simulation experiments with an ad hoc network of 50 mobile nodes are presented in this section. The network, the mobility model, and the propagation model have been described in detail in Section 11.1.4. The setting of the algorithm parameters used in simulations is presented in Table 11.7.

Two network density scenarios are considered. In the low-density scenario, all nodes transmit at 30 mW, and the average node degree in the initial topology is 6.2. In the high-density scenario, the transmit power of each node is 60 mW, and the initial average node degree is 12.13. For each network density, we experiment with two different speed levels ν (4 m/s and 8 m/s). The goal is to study how well the designed algorithm maintains connectivity of the network and how big the backbone is. For this purpose, we calculate the backbone size and compute ratios r_1 and r_2 described in Section 11.1.4.

The simulation results for the low-density scenario are presented in Figure 11.12 and

Table 11.7: Parameters of the algorithm with deterministic pruning

Parameter	Notation	Value
Connectivity probing interval, [sec]	Δ	1
Connectivity information exchange interval, [sec]		1
Maximum age of a database entry, [sec]		1.4
Minimum back-off interval, [sec]	t_{min}	0.4
Maximum back-off interval, [sec]	t_{max}	6
Back-off factor	α	1.5
Initial TTL of a **Probe** message	k	4

Figure 11.13 for average speeds 4 m/s and 8 m/s, respectively. The results for the high-density scenario are presented in Figures 11.14 and 11.15. The first sub-figure in each of the figures demonstrates the two ratios and the second sub-figure shows the number of gateway nodes, or backbone size, over the simulation time. We observe that network connectivity is well maintained, and the backbone recovers fast from occasional disconnections. As expected, the performance in terms of connectivity is better for the dense network. In all four cases the connectivity results are comparable to those obtained by the algorithm with probabilistic pruning (see Section 11.1.4). The backbone size, however, is larger for the algorithm based on deterministic pruning. On the other hand, this algorithm provides a backbone which is more stable, i.e., has smaller fluctuations in size within short time intervals. From the practical point of view, ensuring smooth changes in the backbone should be more appreciated since this gives a more reliable network in fast changing radio propagation environments. From Figures 11.14 and 11.15 we also observe that for the dense network the backbone size tends to decrease over time, indicating that there may be room for more aggressive pruning.

11.3 Discussion and Conclusions

There have been presented two distributed algorithms for asynchronous update of a virtual backbone in mobile networks. With reasonable communication overhead, the algorithms allow for dynamically adapting the backbone to the changing network topology. Our simulation

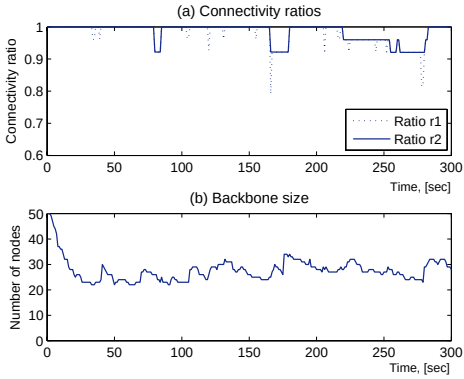


Figure 11.12: Dynamic backbone statistics (low density, $\nu = 4\,\mathrm{m/sec}$).

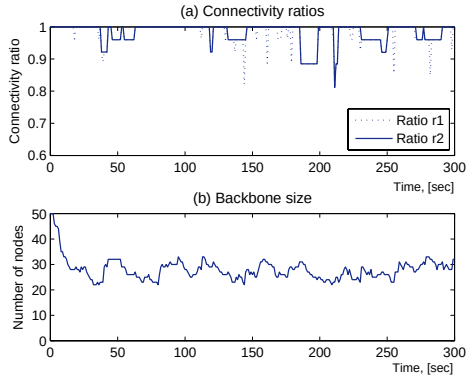


Figure 11.13: Dynamic backbone statistics (low density, $\nu = 8\,\mathrm{m/sec}$).

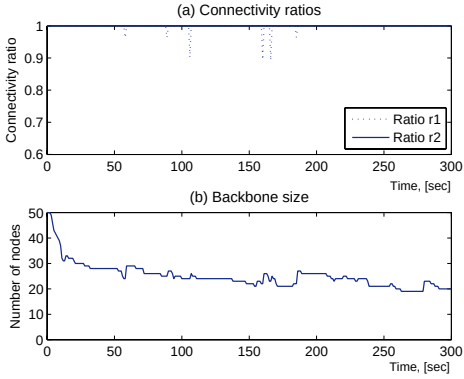


Figure 11.14: Dynamic backbone statistics (high density, $\nu = 4\,\mathrm{m/sec}$).

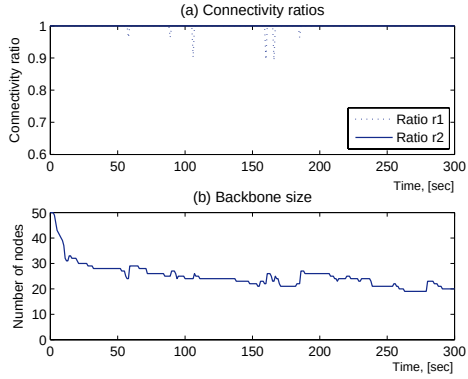


Figure 11.15: Dynamic backbone statistics (high density, $\nu = 8\,\mathrm{m/sec}$).

results demonstrate the effectiveness of the algorithms and their good performance even in sparse networks with high mobility. In particular, time periods where broadcast messages cannot reach all nodes are mostly caused by disconnected physical topology and not backbone management.

Comparing the performance of the two algorithms, we have observed that both algorithms have high reliability in providing very good network connectivity and adapting the virtual backbone to the changing physical topology. The algorithm that uses probabilistic pruning, however, achieves this with a sightly smaller average backbone size. On the other hand, the backbone obtained by the second algorithm is more stable and has significantly smaller fluctuations in size due to more conservative pruning which is an advantage from the practical point of view. Also, the second algorithm is characterized by smaller message overhead.

One of the interesting extensions of the framework is to investigate schemes that allow for dynamic adjustment of the algorithm parameters because this shall make the algorithms more flexible and adaptive, which is in line with the self-configuring nature of ad hoc networks. As a pre-study for this work, more extensive simulations are needed in order to perform a detailed analysis of the effect of algorithm parameters on network connectivity and the backbone size.

Bibliography

[1] I. F. Akyildiz, X. Wangb, and W. Wangb. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, March 2005.

[2] I. F. Akyildiz, S. Weilian, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, Aug. 2002.

[3] K. M. Alzoubi, P.-J. Wan, and O. Frieder. Distributed heuristics for connected dominating sets in wireless ad hocnetworks. *Journal of Communications and Networks*, 4:22–29, 2002.

[4] K. M. Alzoubi, P.-J. Wan, and O. Frieder. Message-optimal connected dominating sets in mobile ad hoc networks. In *Proc. of the Third ACM Intl. Symposium on Mobile Ad Hoc Networking andComputing (MobiHoc '02)*, pages 157–164, June 2002.

[5] K. M. Alzoubi, P.J. Wan, and O. Frieder. New distributed algorithm for connected dominating set in wireless ad hocnetworks. In *Proc. of the 35th Annual Hawaii International Conference on System Sciences(HICSS '02)*, 2002.

[6] C. Ambuhl, A. E. F. Clementi, M. Di Ianni, G. Rossi, and A. Montiand R. Silvestri. The range assignment problem in non-homogeneous static ad-hoc networks. In *Proc. of the 18th International Parallel and Distributed Processing Symposium*, Apr. 2004.

[7] L. Bao and J. J. Garcia-Luna-Aceves. Topology management in ad hoc networks. In *Proc. of the Fourth ACM Intl. Symposium on Mobile Ad Hoc Networkingand Computing (MobiHoc '03)*, pages 129–140, June 2003.

[8] S. Basagni, S. Mastrogiovanni, and C. Petrioli. A performance comparison of protocols for clustering and backbone formationin large scale ad hoc networks. In *Proc. of the first IEEE Intl. Conference on Mobile Ad-hoc and Sensor Systems(MASS '04)*, pages 70–79, Oct. 2004.

[9] J. Blum, M. Ding, A. Thaeler, and X. Cheng. chapter Connected dominating set in sensor networks and MANETs, pages 329–369. Springer US, 2005.

[10] S. Butenko, X. Cheng, C. Oliveria, and P. Pardalos. A new heuristic for the minimum connected dominating set problem on ad hocwireless networks. *Recent Developments in Cooperative Control and Optimization*, pages 61–73, 2004.

[11] M. Čagalj and J.-P. Hubaux. Energy-efficient broadcasting in all-wireless networks. *Journal of Mobile Networks and Applications (MONET)*, pages 177–188, 2003.

[12] J.-H. Chang and L. Tassiulas. Energy conserving routing in wireless ad-hoc networks. In *Proc. of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, pages 21–31, March 2000.

[13] M. Chatterjee, S. Das, and D. Turgut. WCA: A weighted clustering algorithm for mobile ad hoc networks. *Journal of Cluster Computing*, 5:193–204, 2002.

[14] G. Chen, F. G. Nocetti, J. S. Gonzalez, and I. Stojmenovic. Connectivity based k-hop clustering in wireless networks. In *Proc. of the 35th Hawaii International Conference on System Sciences (HICSS'02)*, pages 2450–2459, 2002.

[15] S. Corson and J. Macker. RFC 2501 – Mobile Ad hoc Networking (MANET): Routing protocol performanceissues and evaluation considerations, Jan. 1999.

[16] F. Dai and J. Wu. An extended localized algorithm for connected dominating set formation inad hoc wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 15:908–920, 2004.

[17] A. K. Das, M. El-Sharkawi, R. J. Marks, P. Arabshahi, and A. Gray. Maximization of time-to-first-failure for multicasting in wireless networks:Optimal solution. In *Proc. of IEEE Military Communications Conference (MILCOM '04)*, pages 1358–1363, Oct. 2004.

[18] A. K. Das, R. J. Marks, M. El-Sharkawi, P. Arabshahi, and A. Gray. MDLT: A polynomial time optimal algorithm for maximization of time-to-first-failurein energy constrained wireless broadcast networks. In *Proc. of IEEE GLOBECOM '03*, pages 362–366, 2003.

[19] A. K. Das, R. J. Marks, M. El-Sharkawi, P. Arabshahi, and A. Gray. Minimum power broadcast trees for wireless networks: Integer programmingformulations. In *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, pages 1001–1110, March 2003.

[20] P. Floréen, P. Kaski, J. Kohonen, and P. Orponen. Multicast time maximization in energy constrained wireless networks. In *Proc. of DIALM-POMC Joint Workshop on Foundations of Mobile Computing*, pages 50–58, 2003.

[21] M. R. Garey and D. S. Johnson. *Computers and Intractability: A guide to the theory of NP-completeness*. Series of Books in the Mathematical Sciences. Freeman, 1979.

[22] S. Guha and S. Khuller. Approximation algorithms for connected dominating sets. *Algorithmica*, 20:374–387, 1998.

[23] S. Guha and S. Khuller. Improved methods for approximating node weighted Steiner trees and connecteddominating sets. *Information and Computation*, 150:57–74, 1999.

[24] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. *Mobile Computing*, pages 153–181, 1996.

[25] H. Ju, I. Rubin, K. Ni, and C. Wu. A distributed mobile backbone formation algorithm for wireless ad hoc networks. In *Proc. of the IEEE First Intl. Conference on Broadband Networks (BROADNETS'04)*, pages 661–670, 2004.

[26] I. Kang and R. Poovendran. Maximizing network lifetime of broadcasting over wireless stationary adhoc networks. *Mobile Networks and Applications*, 10:879–896, 2005.

[27] L. M. Kirousis, E. Kranakis, D. Krizanc, and A. Pelc. Power consumption in packet radio networks. *Theoretical Computer Science*, 243(1–2):289–305, July 2000.

[28] D. Kouvatsos and I.-H. Mkwawa. Broadcasting methods in mobile ad hoc networks: An overview. In *Proc. of the Third International Working Conference on Performance Modellingand Evaluation of Heterogeneous Networks (HET-NETs '05)*, July 2005.

[29] X. Li and I. Stojmenović. *Broadcasting and topology control in wireless ad hoc networks*, pages 239–264. CRC Press, 2006.

[30] B. Liang and Z. J. Hass. Virtual backbone generation and maintenance in ad hoc network mobility management. In *Proc. of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, pages 1293–1302, March 2000.

[31] C. R. Lin and M. Gerla. Adaptive clustering for mobile wireless networks. *IEEE Journal of Selected Areas in Communications*, 15(7):1265–1275, Sep. 1997.

[32] H. Liu and R. Gupta. Selective backbone construction for topology control in ad hoc networks. In *Proc. of the first IEEE Intl. Conference on Mobile Ad-hoc and SensorSystems (MASS '04)*, pages 41–59, 2004.

[33] M. Min, X. Huang, S. C.-H. Huang, and W. Wu. Improving construction for connected dominating set with Steiner treein wireless sensor networks. *Journal of Global Optimization*, 35(1):111–119, May 2006.

[34] S.-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu. The broadcast storm problem in a mobile ad hoc network. In *Proc. of the 5th annual ACM/IEEE international conference on Mobile Computingand networking (MobiCom '99)*, pages 151–162, Aug. 1999.

[35] I. Papadimitriou and L. Georgiadis. Minimum-energy broadcasting in wireless networks using a single broadcasttree. *Mobile Networks and Applications*, 11(3):361–375, June 2006.

[36] K. Römer and F. Mattern. The design space of wireless sensor networks. *IEEE Wireless Communications*, 11(6):54–61, Dec. 2004.

[37] P. Santi. Topology control in wireless ad hoc and sensor networks. *ACM Computing Surveys (CSUR)*, 37(2):164–194, June 2005.

[38] P. Sinha, R. Sivakumar, and V. Bharghavan. Enhancing ad hoc routing with dynamic virtual infrastructures. In *Proc. of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2001)*, pages 1763–1772, Apr. 2001.

[39] I. Stojmenović, editor. *Handbook of Sensor Networks: Algorithms and Architectures.* Wiley, Nov. 2005.

[40] I. Stojmenović, M. Seddigh, and J. Zunic. Dominating sets and neighbor elimination-based broadcasting algorithms inwireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 12:14–25, 2001.

[41] I. Stojmenović and J. Wu. Broadcasting and activity-scheduling in ad hoc networks. In S. Basagni, M. Conti, S. Giordano, and I. Stojmenović, editors, *Mobile Ad Hoc Networking*, pages 205–229. IEEE/Wiley, 2004.

[42] S. Venkatesan and C. D. Young. A distributed topology control algorithm for MANETS. In *Proc. of IEEE Military Communications Conference (MILCOM) '05*, 2005.

[43] P.-J. Wan, K. Alzoubi, and O. Frieder. Distributed construction of connected dominating set in wireless ad hocnetworks. In *Proc. of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*, pages 1597–1604, June 2002.

[44] P.-J. Wan, G. Călinescu, X.-Y. Li, and O. Frieder. Minimum-energy broadcast routing in static ad hoc wireless networks. In *Proc. of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2001)*, pages 1162–1171, Apr. 2001.

[45] Y. Wang, W. Wang, and X.-Y Li. Distributed low-cost backbone formation for wireless ad hoc networks. In *Proc. of the 6th ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing(MobiHoc '05)*, pages 2–13, 2005.

[46] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides. On the construction of energy-efficient broadcast and multicast trees inwireless networks. In *Proc. of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, pages 585–594, March 2000.

[47] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides. Distributed algorithms for energy-efficient broadcasting in ad hoc networks. In *Proc. of IEEE Military Communications Conference (MILCOM) '02*, pages 820–825, Oct. 2002.

[48] B. Williams and T. Camp. Comparison of broadcasting techniques for mobile ad hoc networks. In *Proc. of the Third ACM Intl. Symposium on Mobile Ad Hoc Networking andComputing (MobiHoc '02)*, June 2002.

[49] J. Wu and F. Dai. A generic distributed broadcast scheme in ad hoc wireless networks. *IEEE Transactions on Computers*, 53:1343–1354, 2004.

[50] J. Wu and W. Lou. Forward-node-set-based broadcast in clustered mobile ad hoc networks. *Wireless Communications and Mobile Computing*, 3(2):155–173, March 2003.

[51] D. Yuan. Energy-efficient broadcasting in wireless ad hoc networks: Performancebenchmarking and distributed algorithms based on network connectivity characterization. In *Proc. of the 8th ACM/IEEE Intl. Symposium on Modeling, Analysis and Simulationof Wireless and Mobile Systems (MSWiM '05)*, pages 28–35, Oct. 2005.

[52] R. Zheng and R. Kravets. On-demand power management for ad hoc networks. In *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and CommunicationsSocieties (INFOCOM 2003)*, pages 481–491, March 2003.

Appendixes

Appendix A

UMTS Test Networks

Table A.1: General information on test scenarios

Network	Scenario source	Scenario file	Comment
Net1	Ericsson	ericsson.mat	Planning scenario
Net2	ITN	itn72hata.mat	Synthesized scenario
Net3	ITN	itn90hata.mat	Synthesized scenario
Net4	ITN	itn108hata.mat	Synthesized scenario
Net5	ITN	itn129hata.mat	Synthesized scenario
Net6	MOMENTUM	berlin.mat	Planning scenario for the city of Berlin
Net7	MOMENTUM	lisbon1.mat	Planning scenario for the city of Lisbon
Net8	MOMENTUM	lisbon2.mat	Planning scenario for the city of Lisbon

Test networks Net1 and Net6–Net8 are based on realistic planning scenarios that incorporate real terrain and user distribution information as well as calibrated realistic propagation data. The RBS locations, antenna heights, and antenna azimuths are not random but represent a network configuration pre-optimized in an ad hoc manner. In all eight networks, mechanical tilt of all RBS antennas is assumed to be zero. Electrical tilt of six degrees has been applied to all RBS antennas in Net6 and Net7, but it is zero in all other networks. Network Net1 has been provided by Ericsson Research, Linköping. Networks Net6–Net8 have been derived from public scenarios designed within the European project MOMENTUM[1].

Test networks Net2–Net5 are synthesized networks. In these networks, a given number of sites have been randomly (with a uniform distribution) placed over a specified area. For each site, isotropic path-loss predictions have been generated by the Modified Hata model for suburban environment[2] with a Gaussian variation term following zero mean log-normal distribution with standard deviation of 6 dB. Each site has three sectors equipped with directional antennas of type Kathrein 742265 (2140 MHz) with antenna gain of 18 dBi. At each site, the three antenna directions (azimuths) are 0°, 120°, and 240°. Directional losses (gains) have been derived from antenna diagrams interpolated in 3D space assuming a uniform RBS height of 30 m, 1.5 m height for mobile terminals, and zero mechanical and electrical tilt for all antennas.

Tables A.1 and A.2 depict general scenario information and show the test network statistics, respectively. Table A.3 presents the parameter setting used in computational experiments in Part I. In Table A.3, all values are given in linear scale. The orthogonality factor for networks Net1–Net5 is fixed over the entire service area. In networks Net6-Net8, it varies

[1] IST-2000-28088 MOMENTUM (Models and Simulations for Network Planning and Control of UMTS), `http://momentum.zib.de`

[2] ITU-R Report SM 2028-1, Monte-Carlo simulation methodology for the use in sharing and compatibility studies between radio services or systems, 2002.

Table A.2: Network statistics

Network	Total area size (m^2)	Sites	Cells $\lvert\mathcal{I}\rvert$	Grid				Bins covered by one cell** (%)
				Grid size	Total bins	Active bins*, $\lvert\mathcal{J}\rvert$	Bin size (m^2)	
Net1	1280×1800	22	60	32×45	1440	1375	40×40	60.73
Net2	2640×2640	24	72	66×66	4356	4356	40×40	29.15
Net3	4080×4080	30	90	102×102	10404	10404	40×40	32.54
Net4	4560×4560	36	108	114×114	12996	12995	40×40	35.89
Net5	6360×6360	43	129	159×159	25281	25281	40×40	33.60
Net6	7500×7500	50	148	150×150	22500	22500	50×50	26.06
Net7	5000×5000	52	140	250×250	62500	62500	20×20	28.19
Net8	4200×5000	60	164	210×250	52500	52500	20×20	27.31

* *Active bins* are the bins that can be covered by at least one cell.

** In percent to the total number of active bins.

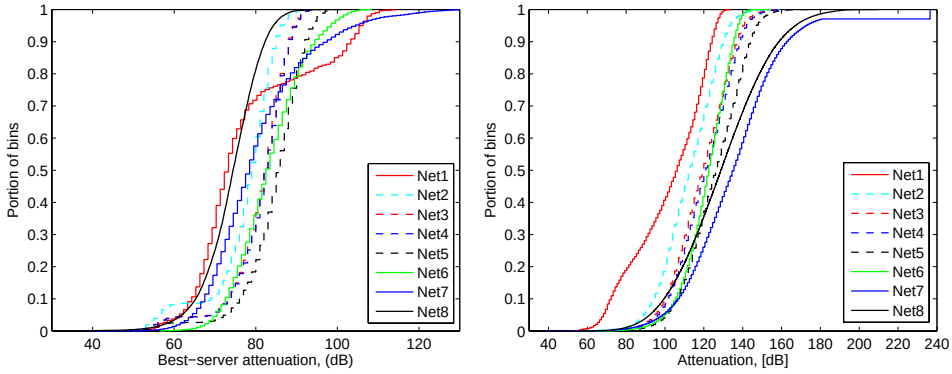
Table A.3: Parameter setting

Network	Min. pilot E_c/I_0 γ_0	Min. pilot RSCP γ_1 (W)	Maximum cell power P_i^{max} (W)	Pilot power bounds (W)		Thermal noise ν_j (W)	Orthogonality factor α_j
				Π_i^{min}	Π_i^{max}		
Net1	0.015	3.16e-15	15	0.2	2.0	1e-13	0.6
Net2-Net5	0.015	3.16e-15	20	0.2	2.6	1e-13	0.6
Net6-Net8	0.010	3.16e-15	20	0.2	2.6	1.55e-14	$\{0.33, 0.63, 0.94\}$

over the network and is equal to 0.33, 0.63, or 0.94, depending on the channel model in bin j (typically urban, mixed, or rural area).

Figures A.1(a) and A.1(b) demonstrate cumulative distributions of the best-server attenuation values and attenuation distribution over the entire areas for all cells in each of the test networks, respectively. For better readability, distribution curves plotted with a solid line correspond to network scenarios based on realistic data (Net1, Net6–Net8), and the other scenarios are represented by dashed lines. Note that in all scenarios attenuation includes also antenna gain.

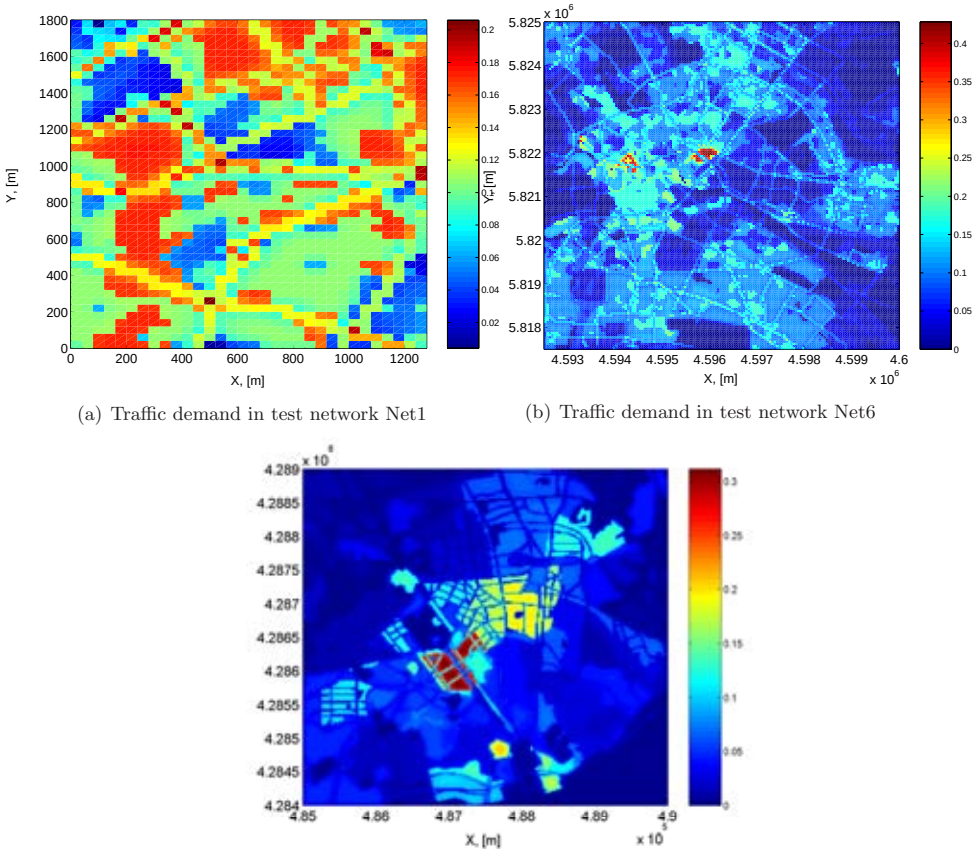
Figures A.2(a), A.2(b), and A.2(c) show distribution of traffic demand for the speech service in Net1, Net6, and Net7, respectively. The traffic demand is given as a static load grid that contains the expected average number of users in each bin at one instance in time.



(a) Best-server attenuation

(b) Attenuation values over the service area

Figure A.1: Cumulative distribution of attenuation values in networks Net1-Net8.



(a) Traffic demand in test network Net1

(b) Traffic demand in test network Net6

(c) Traffic demand in test network Net7

Figure A.2: Traffic distribution in Net1, Net6, and Net7.

Appendix B

The Problem of Minimizing the Number of Cells with Blocked Users

Given a network configuration, user distribution with known traffic demand, and assuming that it is known which user is to be served by which cell, our objective is to minimize the number of cells in which at least one user cannot be served due to limited cell power capacity.

To formulate the problem we use the following set of binary variables (in addition to the set of unknown total transmit power levels $P_i^{Tot}, i \in \mathcal{I}$),

$$x_i = \begin{cases} 1 & \text{if the power capacity of cell } i \text{ is not enough} \\ & \text{to support all users associated with the cell,} \\ 0 & \text{otherwise.} \end{cases}$$

The mathematical formulation of the optimization problem is as follows,

$$\sum_{i \in \mathcal{I}} x_i \longrightarrow \min \tag{B.1a}$$

$$\text{s. t.} \quad (1 - a_{ii}) P_i^{Tot} - \sum_{l \neq i} a_{il} P_l^{Tot} + b x_i \geq c_i \qquad i \in \mathcal{I} \tag{B.1b}$$

$$P_i^{Tot} - x_i P_i^{max} \geq 0 \qquad i \in \mathcal{I} \tag{B.1c}$$

$$P_i^{Tot} \leq P_i^{max} \qquad i \in \mathcal{I} \tag{B.1d}$$

$$x_i \in \{0, 1\} \qquad i \in \mathcal{I} \tag{B.1e}$$

where

$$a_{ii} = \sum_{j \in \bar{\mathcal{J}}_i} \sum_{s \in \mathcal{S}} d_j^s \phi_j^s \cdot (1 - \alpha_j) \ ,$$

$$a_{il} = \sum_{j \in \bar{\mathcal{J}}_i} \sum_{s \in \mathcal{S}} d_j^s \phi_j^s \cdot \frac{g_{lj}}{g_{ij}} \ , \quad \forall l \neq i,$$

$$c_i = \sum_{j \in \bar{\mathcal{J}}_i} \sum_{s \in \mathcal{S}} d_j^s \phi_j^s \cdot \frac{\nu_j}{g_{ij}} + P_i^{CPICH} + P_i^{com} \ .$$

and the other parameters are as used in Section 5.4 in Part I.

In the presented formulation, constraints (B.1b) have been derived after substituting the left-hand side of equation (5.23) by $P_i^{Tot} + b x_i$, where b is a sufficiently big number, and transforming the equality into an inequality. For each cell, the corresponding constraint is ignored if the cell has at least one blocked user (the total DL transmit power of the cell can

be automatically set to its maximum level). Otherwise, the inequality defines the total DL transmit power of the cell. Constraints (B.1c) ensure that if there are some blocked users in a cell, i.e, the total amount of power needed to support all traffic demand exceeds the cell capacity, the total transmit power of the cell is at least at its maximum level P_i^{max}. Constraints (B.1d) set the maximum limit on the total transmit cell power. Thus, by constraints (B.1c) and (B.1d), if $x_i = 1$ for some cell i then the total transmit power of the cell equals its maximum level, i.e., $P_i^{Tot} = P_i^{max}$.

Appendix C

WLAN Test Networks and Parameter Setting

Table C.1: Shadowing propagation model

$$g(\delta, x) = g^{FS}(\delta_0) \cdot \left(\frac{\delta}{\delta_0}\right)^{-\beta} \cdot 10^{0.1 \cdot x}, \quad \text{where} \quad \begin{array}{l} g^{FS}(\delta_0) = \frac{G_{tx} G_{rx} \lambda^2}{(4\pi\delta_0)^2 L} \\ \lambda = \nu / \Upsilon \\ x \in [-\infty, \infty] \\ P(x) = N(\mu, \sigma^2) \end{array}$$

δ	Distance	L	System loss coefficient
δ_0	Reference distance	λ	Wave length
β	Path loss component	ν	Speed of light
G_{tx}	Transmitter antenna gain	Υ	Carrier frequency
G_{rx}	Receiver antenna gain		

Table C.2: Notation and parameter setting

Notation	Parameter	Value
δ_0	Reference distance	1 m
β	Path loss component	5
G_{tx}	1) AP transmitter antenna gain	1.58 (2 dBi)
	2) MT transmitter antenna gain	1.26 (1 dBi)
G_{rx}	AP/MT receiver antenna gain	1 (0 dBi)
L	System loss coefficient	1
ν	Speed of light	$3 \cdot 10^8$ m/s
Υ	Carrier frequency	2.4 GHz
σ	Standard deviation	6 dB
γ^{srv}	Serving threshold	$3.16 \cdot 10^{-12}$ (-90 dBm)
γ^{cs}	CS threshold	10^{-14} (-110 dBm)
$\mathcal{L}$	AP transmit power levels	$\{-1, 4, 7, 10, 13, 15\}$ dBm
L_{MT}	MT transmit power level	31.6 mWatt (15 dBm)