Robust Primary Care Systems

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Martin Comis, M.Sc.

aus Leipzig

Berichter: Univ.-Prof. Dr. rer. nat. Christina Büsing
Univ.-Prof. Dr. rer. pol. Catherine Cleophas
Univ.-Prof. Dr. Ir. Arie M.C.A. Koster

Tag der mündlichen Prüfung: 7. Mai 2021

Martin Comis

Robust Primary Care Systems

D 82 (Diss. RWTH Aachen University, 2021)

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

✉ comis@math2.rwth-aachen.de

Abstract

Primary care systems are generally considered to be the backbone of universal health care. However, as the population ages and the number of primary care physicians declines, this foundation is starting to crumble. There result increasing access distances, waiting times, and workloads up to the point where the system's functioning can no longer be guaranteed. To counteract these developments, representatives from the government, insurances, and associations discuss an array of novel supply concepts and policy changes. This thesis aims to advance this discussion by providing suitable decision support tools, algorithms, and theoretic results. Special attention is thereby put on rural primary care systems, as these are particularly vulnerable due to their geographic-demographic facts. The resulting contributions can be categorized into three main groups and we summarize them hereinafter.

The first part of this thesis addresses the fundamental question of how the quality of primary care systems can be quantified. Due to the inherent complexity and micro-level detail of primary care systems, this turns out to be a highly non-trivial problem and the predominant method of choice is therefore still an assessment of the physician-to-population ratio. To facilitate a more refined analysis, this thesis introduces the hybrid agent-based simulation model SiM-Care. SiM-Care models and tracks the micro-interactions of patients and primary care physicians on an individual level. The model thereby enables decision makers to access several performance indicators such as patient waiting times and physician utilization that can serve as a sound basis for the assessment and comparison of primary care systems. Furthermore, it becomes possible to evaluate changes in the infrastructure, patient behavior, and service design which is impossible with purely ratio-based assessments.

The second part of this thesis examines mobile medical units (MMUs) for the supply of primary care services in rural environments. MMUs are customized vehicles fitted with medical equipment that are easy to relocate and therefore enable a demand-oriented and local provision of health services. Prior to their operation, MMUs necessitate a complex prelaunch strategy to ensure their effectiveness and sustainability. To devise such strategies, this thesis contributes an integrated multi-phased optimization framework. Novel to this framework is the consideration of two types of patient demands, namely patients who seek health services through a centralized appointment system as well as walk-ins who do not announce their visits. Moreover, the framework allows for the incorporation of uncertainties in both types of patient demands which was previously unconsidered.

The third part of this thesis studies two matching problems that derive from the application of MMUs in primary care. It is shown that very restricted variants of these matching problems are already strongly $\mathcal{NP}$-hard. Consequently, this thesis focuses on restricted graph classes and contributes a range of polynomial and pseudo-polynomial algorithms.

Zusammenfassung

Die hausärztliche Versorgung gilt als das Rückgrat der allgemeinen Gesundheitsversorgung. Da die Bevölkerung jedoch altert und die Zahl der Hausärzte zurückgeht, beginnt dieses Fundament zu bröckeln. Die Folge sind zunehmende Anfahrtswege, Wartezeiten und Arbeitsbelastungen bis zu einem Punkt, an dem die Versorgung nicht mehr gewährleistet werden kann. Um diesen Entwicklungen entgegenzuwirken, diskutieren Vertreter von Regierung, Krankenkassen und Verbänden eine Reihe neuer Versorgungskonzepte und politischer Veränderungen. Diese Arbeit möchte durch die Bereitstellung von Entscheidungsunterstützungssystemen, Algorithmen und theoretischen Ergebnissen zu dieser Diskussion beitragen. Besonderes Augenmerk wird dabei auf den ländlichen Raum gelegt, da dieser aufgrund seiner geographisch-demographischen Gegebenheiten besonders anfällig ist. Die entstandenen Resultate lassen sich in drei Hauptgruppen unterteilen und werden nachfolgend diskutiert.

Der erste Teil dieser Arbeit befasst sich mit der grundlegenden Frage, wie die Qualität der hausärztlichen Versorgung quantifiziert werden kann. Da das Gesundheitswesens äußerst komplex ist, erweist sich dies als eine nicht-triviale Fragestellung. Die vorherrschende Methode der Wahl ist daher auch weiterhin die Bewertung des Arzt-Bevölkerungs-Verhältnisses. Um eine verfeinerte Analyse zu ermöglichen, wird in dieser Arbeit das hybride agentenbasierte Simulationsmodell SiM-Care vorgestellt. SiM-Care modelliert die Mikrointeraktionen von Patienten und Hausärzten auf individueller Ebene. Dadurch wird Entscheidungsträgern der Zugang zu mehreren Schlüsselindikatoren wie Patientenwartezeiten und Ärzteauslastung ermöglicht, die als Grundlage zur Bewertung des Versorgungsgrades dienen können. Darüber hinaus ermöglicht es das Modell, Veränderungen der Infrastruktur und des Patientenverhaltens zu analysieren, was mit etablierten Methoden nicht möglich ist.

Der zweite Teil dieser Arbeit untersucht den Einsatz rollender Arztpraxen (MMUs) in ländlichen Gebieten. MMUs sind mit medizinischen Geräten ausgestattete Fahrzeuge, die leicht zu verlegen sind und somit eine wohnortnahe Gesundheitsversorgung ermöglichen. Zur Vorbereitung der Inbetriebnahme von MMUs, muss ein komplexer Planungsprozess durchgeführt werden. Um diesen Planungsprozess zu automatisieren, führt diese Arbeit einen integrierten mehrphasigen Optimierungsansatz ein. Neuartig an diesem Ansatz ist, dass wir zwischen Patienten die ein zentralisiertes Terminsystem verwenden sowie Patienten ohne Termin, sogenannter Laufkundschaft, unterscheiden. Darüber hinaus ermöglicht es der Optimierungsansatz, Unsicherheiten in beiden Patiententypen zu berücksichtigen, was bisher nicht untersucht wurde.

Der dritte Teil dieser Arbeit untersucht zwei Matching Probleme, die aus der Einsatzplanung von MMUs hervorgegangen sind. Es wird gezeigt, dass beide Matching Probleme stark $\mathcal{NP}$-schwer sind. Folglich konzentriert sich diese Arbeit auf eingeschränkte Graphenklassen und entwickelt eine Reihe von polynomiellen und pseudo-polynomiellen Algorithmen.

Acknowledgement

During my time as a Ph.D. student, I received the advice and support of many wonderful individuals towards whom I would like to express my sincere gratitude.

First and foremost, my thanks go to Christina Büsing who has been an exceptional supervisor and whom I could always turn to when I needed advice and guidance. From the beginning, Christina integrated me into the research community and allowed me to attend numerous Ph.D. schools and scientific conferences. I am deeply grateful for these opportunities and I wish her nothing but the best for the future.

The same gratitude and wishes extend to Catherine Cleophas and Arie Koster who not only accepted to act as members of my examination board, but also actively accompanied me through this Ph.D. Their expertise and input has undoubtedly left a mark on this work and Arie deserves a special mention for arranging an unforgettable research stay at the University of Clemson in South Carolina.

Next to my supervisors, I am incredibly thankful for having collaborated with many brilliant minds – regardless of whether our joint efforts made it into this thesis or not. Specifically, these are (in alphabetical order) Mariia Anapolska, Björn Bahl, Timo Gersing, Sebastian Goderbauer, Tabea Krabs, Felix Rauh, Eva Schmidt, Sabrina Schmitz, Manuel Streicher, Felix Willamowski, Sophia Wrede, and Stephan Zieger. Working with them was a great pleasure and proof to me that research should be a team sport. I am equally grateful for the collaborations with the local department of public health, the Actimonda health insurance, and the primary care physicians who hosted me.

The working environment at our chair (which was actually awarded by the university) has always been exceptional and I thank all current and former colleagues for the shared time, thoughts, and lunches. I also had the pleasure of being part of the research training group UnRAVeL which provided a unique research platform for which I thank all members. I feel particularly lucky for having spent the many highs and very few lows at the office with Anto Djoko-Wijono, Timo Gersing, Sascha Kuhnke, and Sabrina Schmitz. Moreover, I would like to thank our current and former IT staff as well as our incredible secretary Angela Hellemeister for their help.

A very special token of appreciation goes to my friends and family who have always been a great moral compass and support. I treasure the many wonderful memories and experiences of the past years and hope that our bonds continue to exist beyond my time here in Aachen. All proofreaders of this thesis deserve my deepest thanks as they caught many of my grammatical and stylistic missteps. In particular I must mention my mother who – being an English

teacher – must have read through hundreds of pages of "dry" Mathematics. Last but not least, I wouldn't be where I am without the unconditional support of my long-term partner Marcia Rückbeil who enriches every day with her encouraging and loving nature – Thank you.

Martin Comis
Aachen, September 20, 2021

Contents

1 Introduction

> *Good health is essential to sustained economic and social development and poverty reduction. Access to needed health services is crucial for maintaining and improving health.*
>
> **— World Health Organization, 2020b**

1.1 Motivation and Research Question

Health is the foundation for the prosperity and well-being of a society (Dodd, 2005). In recognition of this fact, all member states of the World Health Organization have committed themselves to ensure everyone's access to health services without the risk of impoverishment (Dye et al., 2013). Working towards this goal, the majority of member states have established so-called *primary care systems* which have proven to be a reliable and effective mean of achieving universal health coverage (World Health Organization, 2019). Primary care systems "serve as the patient's first point of entry into the health care system and the continuing focal point for all needed health services" (American Academy of Family Physicians, 2019). As such, they are the backbone of efficient, equitable, and resilient health systems (Bitton et al., 2017). The range of primary care services is broad and includes diverse measures such as health promotion, disease prevention, patient education, and counseling as well as the diagnosis and treatment of acute and chronic illnesses (American Academy of Family Physicians, 2019).

Demographic changes in the structure of populations around the world challenge the functioning of today's primary care systems. Medical and technological progress paired with improved living conditions and reduced birth rates have led to an increased share of elderly citizens in virtually every country of the world (United Nations, Department of Economic and Social Affairs, Population Division, 2019). In the United States for example, the percentage of individuals over the age of 65 is predicted to exceed $21\,\%$ of the total population by 2030 (United States Census Bureau, 2017). In Germany, this demographic shift is even more pronounced with citizens aged 65 and above being expected to account for more than a fourth of the total population by 2030; compare Figure 1.1. As populations age, their demand for primary care services tends to increase due to the prevalence of chronic illnesses which disproportionately affect the elderly (Mann et al., 2010; Alemayehu and Warner, 2004). Simultaneously, physicians providing these services are also getting older, e.g., $34.1\,\%$ of all

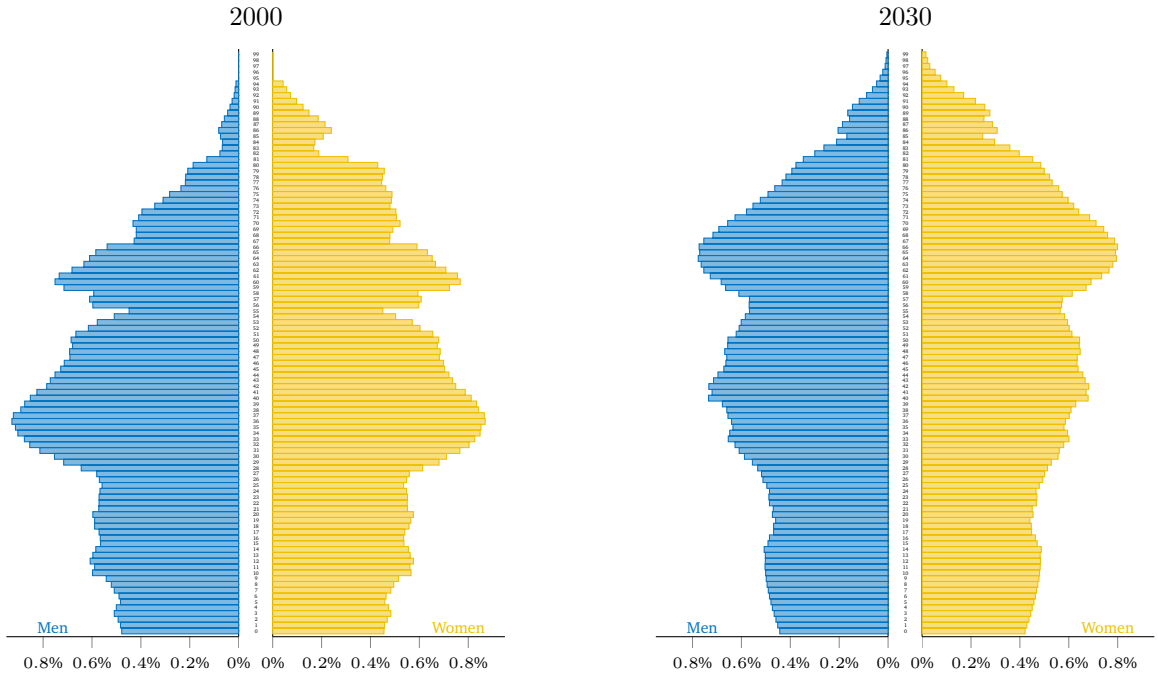


Fig. 1.1.: Age structure in Germany in the year 2000 and projected age structure in the year 2030 under the assumption that birth rates and life expectancy develop moderately and immigration is low (Setting G2-L2-W1) (Bechtold et al., 2019).

primary care physicians in Germany were 60 years or older by the end of 2017 and thus on the verge of retirement (Kassenärztliche Bundesvereinigung, 2017). As fewer medical students are interested in practicing primary care (Mann et al., 2010), let alone willing to invest in a privately-owned primary care practice (Jacob et al., 2015), many retiring physicians will be missing a successor and eventually leave a gap in the primary care supply.

Rural communities, which have always exhibited a substantially lower number of primary care physicians per capita compared to urban areas, are particularly vulnerable to these developments (Mann et al., 2010; Bodenheimer, 1969). The main reason for this is that the low density of health professionals and thus long access distances exacerbate the negative effects that result from the growing imbalance between demand and supply of primary care services (Huigen et al., 1986). However, there are several additional factors that can aggravate the situation in rural settings, e.g., negative migration rates with the younger generation relocating to the urban centers (Huigen et al., 1986) or the absence of adequate public transportation (Murray et al., 1998).

Altogether, this hints at an impending crisis that poses a significant risk of multiplying the existing barriers to rural health services (Mann et al., 2010; Bodenheimer, 1969). In order to manage this crisis and counteract the growing distances between patients and health services, existing primary care systems have to be fundamentally adjusted (Pfaff et al., 2017). However, it is not yet clear what these adjustments might look like. Moreover, all adjustments must be made with the utmost caution, as primary care systems are highly complex and the smallest changes can have disastrous consequences (Fone et al., 2003). This raises a question

of utmost public interest, which emphasizes the special standing of rural primary care in this context and motivates this thesis:

> How can the access to primary care services in rural areas be ensured in light of an aging population and a declining number of primary care physicians?

Questions of such magnitude are intrinsically political and can obviously not be answered by Mathematics. Nevertheless, we are convinced that Mathematics can contribute to the ongoing discussion between statutory health insurances, governments, and the Associations of Statutory Health Insurance Physicians on new concepts and policies that can potentially maintain the standard of health care provision (Rosenbrock and Gerlinger, 2014; Mann et al., 2010). Hence instead of single-handedly taking on the Herculean task of answering the guiding question of this thesis, we develop decision support tools, algorithms, and theory that can hopefully advance the joint efforts to provide solutions. We thereby employ and combine methods from three interrelated fields: simulation sciences, mathematical optimization, and graph theory. The resulting contributions can be structured and grouped into three parts on which we elaborate in the following.

1.2 Contribution of Thesis

In this thesis, we investigate three interrelated subquestions that derive from the guiding question of this thesis. These questions address the quantification of the quality of primary care systems, the operational planning of mobile medical units, and the complexity of specialized matching problems. Accordingly, we structure the contributions of this thesis into three parts and provide an overview of the main results of each part in the following.

Part I: Agent-based Modeling for Primary Care

The planning, analysis, and adaptation of primary care systems is a highly non-trivial problem due to the systems' inherent complexity, unforeseen future events, and scarcity of data. To support the search for solutions, Part I of this thesis introduces the hybrid agent-based simulation model SiM-Care. SiM-Care models and tracks the micro-interactions of patients and primary care physicians on an individual level. At the same time, it models the progression of time via the discrete-event paradigm. Thereby, it enables modelers to analyze multiple performance indicators such as patient waiting times and physician utilization to assess and compare primary care systems. Moreover, SiM-Care can evaluate changes in the infrastructure, patient behavior, and service design. To showcase the strengths of SiM-Care and its validation through expert input and empirical data, we present a case study for a primary care system in the northern Eifel region of Germany. Specifically, we study the immanent implications of demographic change on rural primary care and investigate the effects of an aging population and a decrease in the number of physicians, as well as their combined effects.

Part II: Operational Planning for Mobile Medical Units

Mobile medical units (MMUs) are customized vehicles fitted with medical equipment that are used in the provision of primary care in rural environments. As MMUs can be easily relocated, they enable a demand-oriented, flexible, and local provision of health services. In Part II of this thesis, we investigate the operational planning of an MMU service in three sequential phases to which we refer as Phase 1, Phase 2, and Phase 3.

Phase 1 considers the strategic planning problem for MMUs (SMMU) – a capacitated set covering problem that includes existing practices and two distinct types of patient demands: i) steerable demands representing patients who seek health services through a centralized appointment system and can be steered to any treatment facility within a given consideration set and ii) unsteerable demands representing walk-ins who always visit the closest available treatment facility. We propose an integer linear program for the SMMU that can be solved via Benders decomposition and constraint generation. Starting from this formulation, we focus on the uncertain version of the problem in which steerable and unsteerable demands are modeled as random variables that may vary within a given interval. Using methods from robust optimization and duality theory, we devise exact constraint generation methods to solve the robust counterparts for interval and budgeted uncertainty sets.

In Phase 2, we address the planning of MMUs at the tactical level. To that end, we investigate a bottleneck partitioning variant of the k-center problem that we call the tactical partitioning problem for MMUs (TPMMU). We show that the metric TPMMU is $\mathcal{NP}$-hard to approximate within a constant approximation factor $1 < \alpha < 2$ and subsequently derive a mixed-integer linear programming formulation. Moreover, we show that all our results from Phase 1 for the SMMU translate to a session-specific problem extension that combines strategic and tactical planning and thereby enables for a joint consideration of Phases 1 and 2.

The final Phase 3 is devoted to the vehicle routing of MMUs. For a single depot, we reduce the problem to a minimum weight perfect matching problem in a bipartite graph which can be solved in polynomial time. In the multi-depot setting, we show that the vehicle routing of MMUs is a special case of the so-called budgeted colored bipartite perfect matching problem which we subsequently prove to be strongly $\mathcal{NP}$-hard. To solve the vehicle routing problem for MMUs with multiple depots, we derive a compact integer linear programming formulation.

Last but not least, we evaluate the entire three-phased optimization framework in a computational study based on a set of instances that we generate from the rural primary care system in Germany that we considered in Part I of this thesis.

Part III: Variations of the Matching Problem

Assignment problems are among the most famous combinatorial optimization problems and have been studied in many variations. In Part III of this thesis, we consider two such variations which are motivated by the vehicle routing and staff assignment for MMUs.

The first variation is a weighted matching problem with k independent edge cost functions called the multi-budgeted matching problem (mBM). The total cost of a matching with respect to each cost function must not exceed a corresponding budget. We show that the mBM is strongly $\mathcal{NP}$-hard on paths with uniform edge weights and budgets. Subsequently, we propose a dynamic program for series-parallel graphs with pseudo-polynomial running time for a fixed number of budget constraints. As an extension, we show how this algorithm can be used to solve the mBM on trees using a graph transformation. Realizing that both these graph classes have a bounded treewidth in common, we introduce a dynamic program based on tree decompositions. This approach leads to a pseudo-polynomial algorithm for the mBM with fixed number of budget constraints on graphs of bounded treewidth.

The second matching problem that we study is the minimum color-degree perfect b-matching problem (Col-BM) which represents a new extension of the perfect b-matching problem to edge-colored graphs. The objective of the Col-BM is to minimize the maximum number of differently colored edges in a perfect b-matching that are incident to the same node. We show that the Col-BM is strongly $\mathcal{NP}$-hard on two-colored bipartite graphs and that there exists no α-approximation algorithm for $1 < \alpha < 2$ unless $\mathcal{P} = \mathcal{NP}$. Still, we identify a class of two-colored complete bipartite graphs on which we can solve the Col-BM in polynomial time. Furthermore, we use dynamic programming to devise polynomial-time algorithms solving the Col-BM with a fixed number of colors on series-parallel graphs and simple graphs with bounded treewidth.

1.3 Outline of Thesis

At the beginning of this thesis, we introduce the employed notation and discuss some basic prerequisites (Chapter 2). The main thesis is then structured into three parts: Part I considering the agent-based modeling for primary care, Part II investigating the operational planning for mobile medical units, and Part III studying two variations of the matching problem. Each individual part commences with an introduction that motivates the considered research question in the context of this thesis. We then provide a detailed overview over each part's contributions and discuss the related work. Subsequently follow the technical chapters of each part. As many of these results arose from successful collaborations which have already been published in scientific journals, conference proceedings, and preprints, we inform on the use of previously published materials at the beginning of each part and explicitly name all co-authors in this context. At the end of each part, we conclude and reflect on the presented results before we provide directions for further research.

1.4 Acknowledgment of Funding

The research documented in this thesis has received generous funding from various sources that we would like to acknowledge. This work was supported by the Freigeist-Fellowship of

the Volkswagen Stiftung and the German research council (DFG) Research Training Group 2236 UnRAVeL. We furthermore received funding from the German Federal Ministry of Education and Research (grant no. 05M16PAA) within the project “HealthFaCT - Health: Facility Location, Covering and Transport” as well as congress and traveling grants from the German Academic Exchange Service (DAAD).

2 Preliminaries

The aim of this chapter is to repeat some of the most fundamental prerequisites of this thesis. Our notation and the majority of the topics that we cover here are most likely familiar to the audience of this thesis. We therefore invite the reader to skip this chapter and suggest to come back here whenever a particular notation or concept is unknown. For a comprehensive and much more extensive review of the basics on which this thesis builds, we refer to the excellent textbook by Korte and Vygen (2012).

2.1 General Notation

In this section, we introduce some of the very general notation that is going to be used throughout this thesis.

We denote the set of all integers by $\mathbb{Z}$ and the set of all non-negative integers or natural numbers by $\mathbb{N}$, i.e., we assume that the natural numbers begin by 0. The set of all real numbers will be denoted by $\mathbb{R}$ and we refer to the non-negative real numbers as $\mathbb{R}_{\geq 0}$.

Whenever we refer to sets, we try to use upper case letters while we denote the elements of a set by lowercase letter, e.g., $A = \{a_1, \dots, a_n\}$. We denote the power set of set A by $\mathcal{P}(A) \coloneqq \{A' : A' \subseteq A\}$. The cardinality of set A will be denoted by $|A| \in \mathbb{N}$. Given a mapping $f \colon A \to B$, we will sometimes abbreviate $f(a) =: f_a$ for $a \in A$ to ease notation.

2.2 Graphs and their Properties

Graphs are discrete structures that play a central role in this thesis. In the following, we introduce them along the lines of Korte and Vygen (2012).

An *undirected graph* is an ordered tuple $G = (V, E)$ consisting of a set of *nodes* (or *vertices*) V and a set of *edges* $E \subseteq \{\{v, w\} \ : \ v, w \in V, v \neq w\}$. If the nodes and edges of G are not explicitly defined, we refer to them as $V(G)$ and $E(G)$, respectively. For an edge $e = \{v, w\} \in E$, we call the nodes v and w the *endpoints* of e. Moreover, we say that the nodes v and w are *adjacent* and call w a *neighbor* of v and vice versa. If $v \in V$ is an endpoint of $e \in E$, i.e., $e = \{v, w\}$, we say that e is *incident* to v. Two distinct edges $e_1, e_2 \in E$ are called *adjacent* if they share an endpoint, i.e., $e_1 \cap e_2 \neq \emptyset$. We denote the set of edges that are incident to $v \in V$ as $\delta_G(v) \coloneqq \{e \in E \ : \ v \in e\}$ and write $\delta(v)$ if the corresponding graph G

is clear from context. The number of edges $|\delta(v)|$ incident to $v \in V$ is called the *degree* of v. If $|\delta(v)| = k$ for all $v \in V$, we call the graph G *k-regular*. All $(|V|-1)$-regular graphs in which each pair of nodes is adjacent are called *complete*. For ease of notation, we denote the complete graph with $n = |V|$ nodes by K_n.

A *directed graph* is an ordered tuple $G = (V, A)$ consisting of a set of nodes (or vertices) V and a set of *arcs* (or directed edges) $A \subseteq \{(v, w) \ : \ v, w \in V, v \neq w\}$. In comparison to undirected graphs, the unordered edges are thus replaced by the ordered arcs that have a direction. If the nodes and arcs of G are not explicitly defined, we refer to them as $V(G)$ and $A(G)$, respectively. For an arc $a = (v, w) \in A$, we call the node v the *tail* and the node w the *head* of a. Moreover, we say that arc a *leaves* v and *enters* w. We denote the set of all arcs that leave $v \in V$ as $\delta^+_G(v) \coloneqq \{(v, w) \in A\}$ and write $\delta^+(v)$ if the corresponding graph G is clear from context. Analogously, we denote the set of all arcs that enter $v \in V$ as $\delta^-_G(v) \coloneqq \{(w, v) \in A\}$ and shortly write $\delta^-(v)$. The number of leaving arcs $|\delta^+(v)|$ is called the *out-degree* of $v \in V$ and the number of entering arcs $|\delta^-(v)|$ is called the *in-degree* of $v \in V$. For a set of nodes $S \subseteq V$, we denote the set of all leaving arcs as $\delta^+(S) \coloneqq \{(v, w) \in A : v \in S, w \in V \setminus S\}$ and the set of all entering arcs as $\delta^-(S) \coloneqq \{(w, v) \in A : v \in S, w \in V \setminus S\}$.

When we refer to a *graph*, we either mean a directed or undirected graph. Graphs, as we defined them above, are also called *simple* graphs to distinguish them from so-called *multi*-graphs which may contain multiple copies of edges and so-called *loops* in which both endpoints coincide. If not stated otherwise, we generally assume graphs to be simple.

For an undirected (directed) graph G, a sequence $W = v_1, e_1, v_2, \dots, v_k, e_k, v_{k+1}$ with $k \geq 0$ and $\{v_i, v_{i+1}\} \in E(G)$ ($(v_i, v_{i+1}) \in A(G)$) for all $i \in \{1, \dots, k\}$ is called an *edge progression* in G. If we additionally have that $e_i \neq e_j$ for $i \neq j$, we call W a *walk* in G. The sequence W is called *closed* if $v_1 = v_{k+1}$. The graph $P = (\{v_1, \dots, v_{k+1}\}, \{e_1, \dots, e_k\})$ for a given walk $W = v_1, e_1, v_2, \dots, v_k, e_k, v_{k+1}$ with $v_i \neq v_j$ for $i \neq j$ is called a *path* from v_1 to v_{k+1} of length k. The graph $C = (\{v_1, \dots, v_k\}, \{e_1, \dots, e_k\})$ for a given closed walk $W = v_1, e_1, v_2, \dots, v_k, e_k, v_1$ with $v_i \neq v_j$ for $i \neq j$ is called a *cycle* in G of length k.

An undirected graph G is called *connected* if there exists a path between any two nodes $v, w \in V(G)$ with $v \neq w$; otherwise we call G *disconnected*.

An undirected graph G is called a *forest* if it does not contain a cycle. If a forest is additionally connected, we call it a *tree*. The nodes of a tree with degree one are called *leaves*. A tree $T = (V, E)$ with a designated node $r \in V$ is known as a *rooted tree* and we call r the *root* of T. The edges of a rooted tree can be naturally oriented away from the root. Based on this natural order, we call the successors of a node $v \in V$ the *children* of v and the unique predecessor of v the *parent* of v. In general, such directed trees are also known as *arborescences*, however we will not cover them in the extent of this thesis.

If the nodes of an undirected graph $G = (V, E)$ can be partitioned into two sets $V_A, V_B \subseteq V$ such that $V_A \cap V_B = \emptyset$, $V_A \cup V_B = V$, and $E \subseteq \{\{v, w\} : v \in V_A, w \in V_B\}$, we call G *bipartite* and write $G = (V_A \cup V_B, E)$. In a bipartite graph, all edges have one endpoint in V_A and one

endpoint in V_B. If all nodes in V_A are pairwise adjacent to all nodes in V_B, we say that G is *completely* bipartite. We denote the complete bipartite graph with $m = |V_A|$ and $n = |V_B|$ nodes by $K_{m,n}$.

Finally, we define a *weighted* undirected (directed) graph as a pair of an undirected (directed) graph G and a function $w : E(G) \to \mathbb{N}$ ($w : A(G) \to \mathbb{N}$) that assigns a weight $w(e)$ to all edges of G.

2.3 Matchings

Matchings define a simple structure on graphs and we cover them in this separate section due to their significance in this thesis.

Let $G = (V, E)$ be an undirected graph. We call a set of edges $M \subseteq E$ a *matching* if all edges $e \in M$ are pairwise disjoint. Hence, if we denote the set of edges in an edge subset $M \subseteq E$ that are incident to $v \in V$ as $\delta_M(v) \coloneqq \delta(v) \cap M$, the edge subset M is a matching if and only if $|\delta_M(v)| \leq 1$ for all $v \in V$. For a matching $M \subseteq E$ and node $v \in V$, we say that v is *matched* by M if $|\delta_M(v)| = 1$; otherwise call v *unmatched*. We call $|M|$ the *size* of the matching M and if $|M| \geq |M'|$ holds for all matchings $M' \subseteq E$, we call M a *maximum* matching. In the special case that M matches all nodes in V, i.e., $|M| = |V|/2$, we call M a *perfect* matching.

A generalization of matchings to undirected graphs $G = (V, E)$ with a *balance function* $b\colon V \to \mathbb{N}$ are known as *b-matchings*. A b-matching is an edge subset $M \subseteq E$ such that each node $v \in V$ is incident to at most $b(v)$ edges in M, i.e., $|\delta_M(v)| \leq b(v)$. As a result, matchings are special b-matchings with $b(v) = 1$ for all $v \in V$. We call $|M|$ the *size* of the b-matching M and if $|M| \geq |M'|$ holds for all b-matchings $M' \subseteq E$, we call M a *maximum* b-matching. In the special case that $|\delta_M(v)| = b(v)$ for all $v \in V$, we refer to M as a *perfect* b-matching.

Both matchings and b-matchings can be extended to weighted graphs and are well studied. A summary of the most important results can be found in Korte and Vygen (2012).

2.4 Complexity Theory

Complexity theory is a highly technical field that revolves around alphabets, languages, and (non-)deterministic Turing machines. As such a deep understanding is not necessary in this thesis, we try to keep it as simple as possible. The interested reader is referred to Garey and Johnson (1979) for a more formal take at the topic. This brief summary of the most basic concepts is in structure and content closely related to Korte and Vygen (2012).

The elementary problems studied in complexity theory are so-called decision problems – problems which only allow for a Yes or No answer.

Definition 2.1. A *decision problem* is a pair $\Pi = (\mathcal{I}, \mathcal{J})$ where the elements of $\mathcal{I}$ are called the *instances* of Π, the elements of $\mathcal{J} \subseteq \mathcal{I}$ are called the Yes-*instances*, and the elements of $\mathcal{I} \setminus \mathcal{J}$ are called the No-*instances*. The decision problem Π asks for given $I \in \mathcal{I}$ whether I is a Yes-instance, i.e., whether $I \in \mathcal{J}$.

To decide or solve a decision problem, we generally look for an algorithm that can determine whether a given instance is a Yes-instance or not. Formally, this leads to the following definition.

Definition 2.2. An *algorithm* for the decision problem $\Pi = (\mathcal{I}, \mathcal{J})$ is a function $\mathcal{A}\colon \mathcal{I} \to \{0, 1\}$ that is defined by $\mathcal{A}(I) = 1$ if $I \in \mathcal{J}$ and $\mathcal{A}(I) = 0$ if $I \in \mathcal{I} \setminus \mathcal{J}$.

A key property of an algorithm is the number of elementary steps (e.g., assignments, arithmetic operations, or comparisons) required to compute $\mathcal{A}(I)$ which depends on the given instance $I \in \mathcal{I}$, in particular its "size".

Instances $I \in \mathcal{I}$ usually consist of a list of numbers that are encoded as binary strings. The length of this binary encoding, i.e, the number of bits required to represent I, is known as the *encoding size* of I that we denote by $\operatorname{size}(I) \in \mathbb{N}$. For example, an integer $n \in \mathbb{Z}$ has encoding size $\operatorname{size}(n) = \lfloor \log_2(|n|) \rfloor + 2$ if we use a binary representation plus an extra bit to represent the sign of n. Encodings are generally not unique, e.g., integers are usually represented using an alternative encoding known as *two's complements*; compare Java (Oracle, 2018). In the following, we will not explicitly discuss the employed encoding but simply assume that it is *efficient* which means that its length is polynomially bounded by the minimum possible encoding length. We can now define the running time of an algorithm.

Definition 2.3. Let $\mathcal{A}\colon \mathcal{I} \to \mathbb{Z}$ be an algorithm and let $f\colon \mathbb{N} \to \mathbb{N}$ be a computable function. If there exists $c \in \mathbb{N}$ such that $\mathcal{A}$ terminates in at most $c \cdot f(\operatorname{size}(I))$ elementary steps for each instance $I \in \mathcal{I}$, we say that $\mathcal{A}$ *runs* in $\mathcal{O}(f(\operatorname{size}(I)))$ time. Also, we say that the *running time* of $\mathcal{A}$ is in $\mathcal{O}(f(\operatorname{size}(I)))$.

The running time of an algorithm is a measure of how long it will (in the worst case) take to decide instance $I \in \mathcal{I}$ using algorithm $\mathcal{A}$. Obviously, we are mostly interested in fast and efficient algorithms that we define as follows.

Definition 2.4. An algorithm $\mathcal{A}\colon \mathcal{I} \to \mathbb{Z}$ runs in *polynomial time*, if there exists a polynomial $p\colon \mathbb{N} \to \mathbb{N}$ such that $\mathcal{A}$ runs in $\mathcal{O}(p(\operatorname{size}(I)))$ time for all $I \in \mathcal{I}$. We then say that $\mathcal{A}$ is a *polynomial-time* algorithm.

Polynomial algorithms are the ones that *efficiently* decide decision problems. As a result, we define the set of all decision problems for which such a polynomial-time algorithm exists.

Definition 2.5. The class $\mathcal{P}$ consists of all decision problems Π for which a polynomial-time algorithm exists.

We will now introduce another class of decision problems for which we no longer require a polynomial-time algorithm, but instead the existence of a so-called *certificate* that allows us to decide all Yes-instances in polynomial time.

Definition 2.6. The class $\mathcal{NP}$ consists of all decision problems $\Pi = (\mathcal{I}, \mathcal{J})$ such that for each Yes-instance $I \in \mathcal{J}$ there exists a *certificate* $c(I) \in \{0,1\}^{\text{size}(I)}$ such that I can be decided in polynomial time when $c(I)$ is given.

As all instances of a decision problem $\Pi \in \mathcal{P}$ can be decided in polynomial time, it obviously follows that $\mathcal{P} \subseteq \mathcal{NP}$. However, whether $\mathcal{P} = \mathcal{NP}$ or if $\mathcal{P} \neq \mathcal{NP}$ is the longstanding open question that is central to complexity theory and worth one million dollars to the one solving it (Devlin, 2002). In this thesis, we work under the assumption that $\mathcal{P} \neq \mathcal{NP}$, however we try to repeat this assumption wherever it is required or implied.

For many decision problems in $\mathcal{NP}$ it is unknown whether they have a polynomial-time algorithm. In the following, we introduce a concept that allows us to conclude that some problems in $\mathcal{NP}$ are at least as hard as all other problems in $\mathcal{NP}$. The foundation of this concept are the so-called *polynomial reductions*.

Definition 2.7. Let $\Pi_1 = (\mathcal{I}_1, \mathcal{J}_1)$ and $\Pi_2 = (\mathcal{I}_2, \mathcal{J}_2)$ be two decision problems. We say that Π_1 *polynomially reduces* to Π_2 is there exists a function $f \colon \mathcal{I}_1 \to \mathcal{I}_2$ that can be computed in polynomial time such that $f(I) \in \mathcal{J}_2$ for all $I \in \mathcal{J}_1$ and $f(I) \in \mathcal{I}_2 \setminus \mathcal{J}_2$ for all $I \in \mathcal{I}_1 \setminus \mathcal{J}_1$.

The key observation to polynomial reductions is the following.

Corollary 2.8. *If Π_1 polynomially reduces to Π_2, every polynomial-time algorithm for Π_2 implies a polynomial-time algorithm for Π_1.*

We can now define the class of all decisions problems that are at least as hard as all other problems in $\mathcal{NP}$.

Definition 2.9. A decision problem $\Pi \in \mathcal{NP}$ is called *$\mathcal{NP}$-complete* if all problems in $\mathcal{NP}$ polynomially reduce to Π.

As a result of Corollary 2.8, the existence of a polynomial-time algorithm for any $\mathcal{NP}$-complete problem implies that $\mathcal{P} = \mathcal{NP}$. This motivates us to perceive $\mathcal{NP}$-complete problems as being the "hard" problems opposed the "easy" problems in $\mathcal{P}$.

Fundamental to the concept of $\mathcal{NP}$-completeness is obviously that there are $\mathcal{NP}$-complete problems and indeed, Cook (1971) showed that the well-known satisfiability problem SAT is $\mathcal{NP}$-complete.

2.5 Parameterized Complexity Theory

Parameterized complexity theory is an extension of the previously introduced complexity theory. In this section, we provide a brief introduction into this extension that is mostly based on Koster (2017).

Parameterized complexity theory studies so-called *parameterized problems* that extend the previously introduced decision problems by a *parameter*.

Definition 2.10. A *parameterized problem* is a pair (Π, κ), where $\Pi = (\mathcal{I}, \mathcal{J})$ is a decision problem and $\kappa : \mathcal{I} \to \mathbb{N}$ is a polynomial-time computable function called *parameter*.

A core concept in study of parameterized problems is the so-called *fixed parameter tractability* that requires a polynomial-time algorithm in the input size but allows for a non-linear factor that depends on the parameter.

Definition 2.11. Let (Π, κ) be a parameterized problem with $\Pi = (\mathcal{I}, \mathcal{J})$. An algorithm $\mathcal{A}$ is called *fixed parameter tractable* $(\mathcal{FPT})$ with respect to parameter κ, if there exists a computable function $f\colon \mathbb{N} \to \mathbb{N}$ and a polynomial $p\colon \mathbb{N} \to \mathbb{N}$ such that for every instance $I \in \mathcal{I}$ the running time of $\mathcal{A}$ is in $\mathcal{O}(f(\kappa(I)) \cdot p(\text{size}(I)))$. We call a parameterized problem (Π, κ) *fixed parameter tractable* $(\mathcal{FPT})$ if there exits an $\mathcal{FPT}$-algorithm with respect to κ.

It is crucial to the definition of fixed parameter tractability that the non-linear term in the algorithm's running time is independent of the encoding size. In the following, we relax this requirement and allow for polynomials in the encoding size for which the degree may depend on the parameter.

Definition 2.12. Let (Π, κ) be a parameterized problem with $\Pi = (\mathcal{I}, \mathcal{J})$. An algorithm $\mathcal{A}$ is called an *$\mathcal{XP}$-algorithm* with respect to parameter κ, if there exists a computable function $f\colon \mathbb{N} \to \mathbb{N}$ such that for every instance $I \in \mathcal{I}$, the running time of $\mathcal{A}$ is in $\mathcal{O}(\text{size}(I)^{f(\kappa(I))})$. We call a parameterized problem (Π, κ) $\mathcal{XP}$ if there exits an $\mathcal{XP}$-algorithm with respect to κ.

Clearly, if a parameterized problem is $\mathcal{FPT}$ it is also $\mathcal{XP}$. A more refined analysis of parameterized problems can be achieved through the so-called W hierarchy (Downey and Fellows, 1999) on which we will not elaborate as this is beyond the scope of this thesis.

2.6 $\mathcal{NP}$-hardness

Up to this point, we exclusively considered the complexity analysis of decision problems. However, most problems in combinatorial optimization do not seek for a Yes/No answer, but instead a feasible solution that optimizes a given objective. As we are going to see in this

section, most of the previously described concepts transfer to optimization problems. The following outline of the resulting formalisms is along the lines of Korte and Vygen (2012).

To begin with, we formally define what we understand by an optimization problem.

Definition 2.13. An *optimization problem* is a tuple $\Pi = (\mathcal{I}, (\mathcal{S}_I)_{I\in\mathcal{I}}, c, \mathrm{goal})$, where $\mathcal{I}$ is the set of *instances*; $\mathcal{S}_I$ are the *feasible solutions* for instance $I \in \mathcal{I}$; $c(I, x) \in \mathbb{Z}$ is the *solution value* of $x \in \mathcal{S}_I$ for $I \in \mathcal{I}$; and $\mathrm{goal} \in \{\min, \max\}$. We define the *optimal solution value* of an instance $I \in \mathcal{I}$ as $\mathrm{OPT}(I) \coloneqq \mathrm{goal}_{x\in\mathcal{S}_I}\ c(I, x)$.

An (exact) *algorithm* $\mathcal{A}$ for an optimization problem $\Pi = (\mathcal{I}, (\mathcal{S}_I)_{I\in\mathcal{I}}, c, \mathrm{goal})$ computes for every instance $I \in \mathcal{I}$ with $\mathcal{S}_I \neq \emptyset$ an *optimal solution* $\mathcal{A}(I) \in \mathcal{S}_I$ with $c(I, \mathcal{A}(I)) = \mathrm{OPT}(I)$. Polynomial reductions naturally extend to optimization problems and we define the following class of problems.

Definition 2.14. An optimization or decision problem Π is called $\mathcal{NP}$-hard if all problems in $\mathcal{NP}$ polynomially reduce to Π.

By definition, all $\mathcal{NP}$-complete decision problems are also $\mathcal{NP}$-hard.

Next, we introduce a concept that allows us to classify $\mathcal{NP}$-hard optimization or decision problem even further. To that end, recall that a unary encoding represents instances $I \in \mathcal{I}$ as strings composed of a single symbol (often denoted by 1) and blanks. Unary encodings are usually much longer binary encoding. For example, a natural number $n \in \mathbb{N}$ is represented by n repetitions of the symbol $(11\ldots1)$. Hence, $\mathrm{size}(n) = n$ in unary encoding. By using a unary encoding of instances, we artificially reduce the running time requirements of polynomial-time algorithms. This gives rise to the following class of algorithms.

Definition 2.15. Let Π be a decision or optimization problem. An algorithm $\mathcal{A}$ for Π is said to run in *pseudo-polynomial time*, if it runs in polynomial time when we use a unary encoding of the instances $\mathcal{I}$. We then say that $\mathcal{A}$ is a *pseudo-polynomial* algorithm.

Pseudo-polynomial algorithms are generally not polynomial-time algorithms, as their running times can be exponential in the encoding size with respect to an efficient binary encoding. Nevertheless, pseudo-polynomial algorithms may be useful in practice which gives rise to the following definition.

Definition 2.16. An $\mathcal{NP}$-hard optimization or decision problem Π is called *weakly* $\mathcal{NP}$-hard if it has a pseudo-polynomial algorithm.

Given the existence of a pseudo-polynomial algorithm, weakly $\mathcal{NP}$-hard problems can be considered as the "easier" $\mathcal{NP}$-hard problems. As a counterpart, we consider the so-called *strongly* $\mathcal{NP}$-hard problems.

Definition 2.17. An optimization or decision problem Π is called *strongly* $\mathcal{NP}$-hard if it remains $\mathcal{NP}$-hard even if we use a unary encoding of the instances $\mathcal{I}$.

By definition, strongly $\mathcal{NP}$-hard problems cannot have a pseudo-polynomial algorithm unless $\mathcal{P} = \mathcal{NP}$. In that sense they are "harder" than the weakly $\mathcal{NP}$-hard problems.

Parameterized complexity theory can also be extended to optimization problems, however we refrain from doing within the scope of this thesis.

2.7 Approximation Algorithms

Approximation algorithms are polynomial-time algorithms that are not exact but instead only ensure a relative performance guarantee, i.e., the obtained solution value may differ from the optimal solution value by at most a multiplicative factor. The following formalization of the concept of approximation algorithms is based on the one in Korte and Vygen (2012).

The notion of approximation algorithms is clearly only meaningful for optimization problems. Therefore, let $\Pi = (\mathcal{I}, (\mathcal{S}_I)_{I \in \mathcal{I}}, c, \mathrm{goal})$ be an optimization problem with non-negative solution values, i.e., we assume in the following that $c(I, x) \geq 0$ for all $I \in \mathcal{I}$ and all $x \in \mathcal{S}_I$. As before, we denote the optimal solution value of instance $I \in \mathcal{I}$ by $\mathrm{OPT}(I)$. Moreover, we denote the solution value for instance $I \in \mathcal{I}$ obtained by algorithm $\mathcal{A}$ as $\mathcal{A}(I) \coloneqq c(I, \mathcal{A}(I))$. We can now formalize an approximation algorithm.

Definition 2.18. Let $\Pi = (\mathcal{I}, (\mathcal{S}_I)_{I \in \mathcal{I}}, c, \mathrm{goal})$ be an optimization problem with non-negative solution values and $\alpha > 1$. An *α-approximation algorithm* for Π is a polynomial-time algorithm $\mathcal{A}$ for Π such that

$$\frac{1}{\alpha} \mathrm{OPT}(I) \leq \mathcal{A}(I) \leq \alpha \, \mathrm{OPT}(I) \tag{2.1}$$

for all instances $I \in \mathcal{I}$. We call $\alpha > 1$ the *performance ratio* or *performance guarantee* of $\mathcal{A}$.

The first inequality in (2.1) applies if Π is a maximization problems ($\mathrm{goal} = \max$), while the second applies when Π is a minimization problem ($\mathrm{goal} = \min$). For all instances $I \in \mathcal{I}$ with $\mathrm{OPT}(I) = 0$, the definition above requires an optimal solution, i.e., it must hold that $\mathcal{A}(I) = 0$.

Note, that our definition of approximation algorithms excludes 1-approximation algorithms as these are the exact polynomial-time algorithms of an optimization problem Π.

Part I

Agent-based Modeling for Primary Care

The Decision Support Tool SiM-Care

3 Introductory Remarks and Contribution

> *Primary health care has been proven to be a highly effective and efficient way to address the main causes and risks of poor health and well-being today, as well as handling the emerging challenges that threaten health and well-being tomorrow.*
>
> **— World Health Organization, 2020a**

3.1 Motivation and Research Question

The simultaneous rising of the demand for primary care services and decreasing supply of health professionals require fundamental adjustments of primary care systems (Pfaff et al., 2017). Statutory health insurances, governments, and the Associations of Statutory Health Insurance Physicians therefore discuss and explore a variety of new concepts and policies to uphold the current standard of health care provision (Rosenbrock and Gerlinger, 2014; Mann et al., 2010). These adjustments are commonly classified into i) microsystem improvements, which aim at enhancing a single server of the system and can be implemented at an individual level, and ii) macrosystem reforms, which are fundamental system-wide changes that affect all servers and must be implemented by policy makers (Zhong et al., 2016). In the context of primary care, microsystem improvements focus on single practices and consider changes that can be implemented independently by primary care physicians (PCPs), e.g., new appointment scheduling strategies or work-flows. Macrosystem reforms on the other hand, are more holistic and consider, e.g., the distribution of PCPs, compensation schemes, or patient transportation.

What is common to both types of system changes, is that they need to be validated and evaluated prior to their potentially costly implementation (Pfaff et al., 2017). Naturally, this leads to the following pressing question:

> How can the quality of primary care systems and the effects of changes made to them be quantified?

In German legislation, this question is addressed by the 2012 GKV-Versorgungsstrukturgesetz (Gemeinsamer Bundesausschuss, 2012), which defines adequate health care supply on the

basis of profession-specific ratios. The law subdivides Germany into zones and specifies the required population-to-provider ratio for each medical specialization. For example, the predefined nominal ratio of primary care physicians is one PCP per $1{,}609$ inhabitants (Gemeinsamer Bundesausschuss, 2012, §11(4)). This base indicator can be adjusted to account for a zone's individual demographic and geographic characteristics (Gemeinsamer Bundesausschuss, 2012, §2). If the actual ratio of a zone is significantly higher than the nominal ratio, closing practices will not be replaced. If it is significantly lower, new practices are permitted to be opened.

Beyond Germany, we can find similar ratio-based measures in other European countries like Bulgaria, Estonia, Italy, and Spain (Kringos et al., 2015). In the United States, the Health Resources and Services Administration (HRSA) defines adequate health care supply based on profession- and region-specific population-to-provider ratios. If the predefined population-to-provider ratio of a geographic area is exceeded, HRSA designates it a health professional shortage area to which National Health Service Corps personnel is directed with priority. Specifically, for primary care this predefined ratio is $3{,}500$ to 1 which can be lowered to $3{,}000$ to 1 if a region's needs are unusually high (Bureau of Health Workforce, Health Resources and Services Administration (HRSA), 2020).

Obviously all these ratio-based assessments have several shortcomings. Even after adjustment, population-to-provider ratios can only provide a very rough estimate of the actual demand for health services. Furthermore, adjustment criteria are highly dependent on the definition of the underlying zones or geographic areas. Factors such as the accessibility of practices and the PCPs' individual workloads are completely neglected. Finally, ratio-based assessments cannot account for new health care delivery concepts such as telemedicine, mobile medical units, or centralized appointment scheduling as these do not affect the evaluated population-to-provider ratios.

To overcome these limitations, a new approach to model the dynamic effects in primary care systems is required. However, analyzing and evaluating health care systems is a complex and complicated task due to the large number of involved individuals and uncertain nature of health care processes, e.g., fluctuating demands, arrival times of patients, emergency patients, and durations of treatments. There result so-called *wicked* problems (Rittel and Webber, 1973); problems that stem from a "world of diverse, pluralistic and dynamic changes that is ill-suited to traditional optimization and equilibrium modeling" (Zellner and Campbell, 2015). As agent-based modeling (ABM) can account for individual agents and their interactions on the micro-level, some consider it a promising paradigm to model the type of complex systems that cause wicked problems (Zellner and Campbell, 2015). A general introduction to the concept of ABM is provided in Gilbert (2008). Existing studies implementing ABM have considered diverse social systems. Examples include matters such as tobacco control (Rigotti and Wallace, 2015) and educational policy research (Maroulis et al., 2010). There are also numerous applications of ABM in the field of health care (Barnes et al., 2013), e.g., for accountable care organizations (Alibrahim and Wu, 2018; Liu and Wu, 2016), for medical

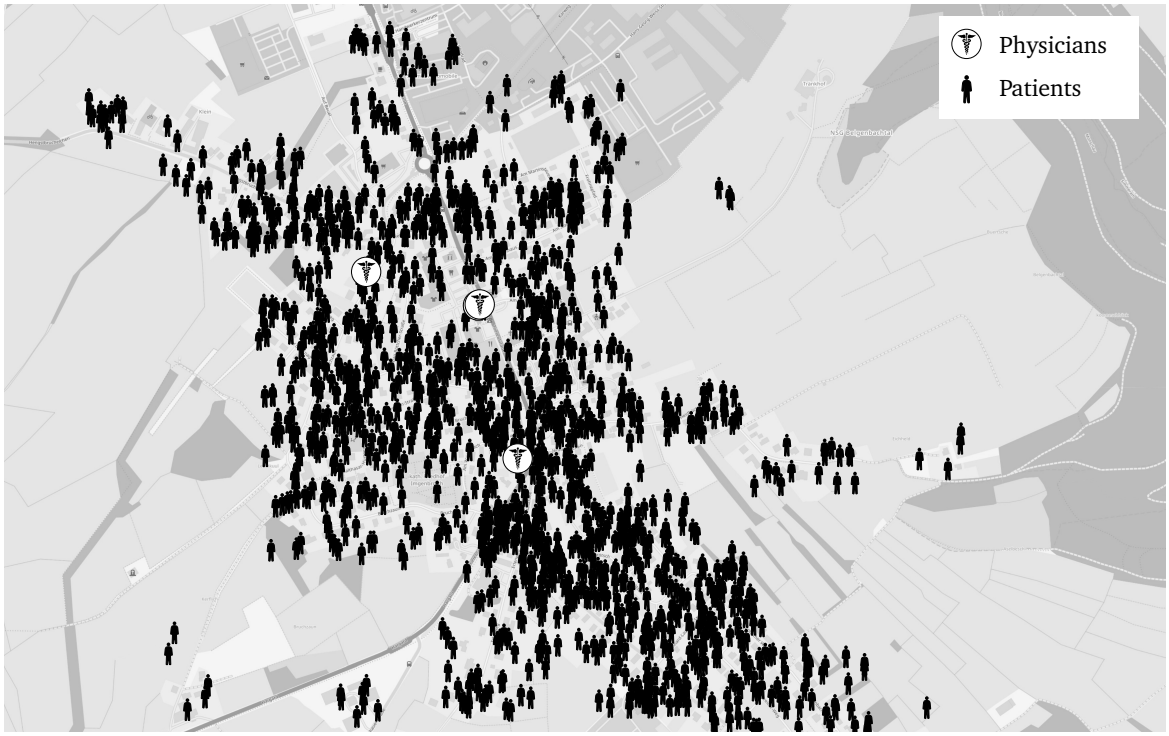


Fig. 3.1.: Geo-social system of patients and physicians.[1]

workforce forecasting (Lopes et al., 2018), and in epidemiology (Patlolla et al., 2006; Meng et al., 2010).

In the following, we employ the concept of agent-based modeling to develop a novel simulation tool that is designed to support the assessment of primary care systems.

SiM-Care – A Simulation Model for Primary Care

This thesis introduces the hybrid agent-based simulation tool SiM-Care (**Si**mulation **M**odel for Primary **Care**) to model the dynamics of primary care systems. SiM-Care models patients and PCPs on an individual level as illustrated by Figure 3.1: Patients and PCPs are modeled via a geo-social system in which patients decide whether and where to request an appointment based on their preferences and state of health; and PCPs handle appointment requests, manage patient admissions, and treat patients. By tracking the resulting interactions in SiM-Care, planners can identify dependencies of different subproblems, evaluate new planning approaches, and quantify the effects of interventions on the basis of multiple meaningful performance measures. From empirical data, we develop realistic test scenarios including a controllable degree of uncertainty realized via stochastic simulation experiments.

Before we discuss existing work in the field of decision support for health care planning that is related to SiM-Care, we summarize our contributions.

[1] Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

3.2 Contribution

The main contribution of Part I of this thesis can be summarized as follows. We introduce the simulation model SiM-Care, which provides decision makers with a versatile decision support tool for primary care planning. SiM-Care is a generic model that can be easily modified and extended to meet each modeler's needs. Patients and physicians are modeled as individuals who follow their own objectives, learn, and adjust. To ensure computational tractability, the model incorporates a global event queue at its core. As such, SiM-Care can be considered as an integrated hybrid simulation model that combines paradigms from agent-based modeling and discrete event simulation. Based on empirical data from a German primary care system, we illustrate how scenarios for the simulation model can be generated. Finally, we showcase the opportunities of SiM-Care through a case study and perform a sensitivity analysis. To the best of our knowledge, SiM-Care is the first model of its kind that captures entire primary care systems with all physicians and patients as individual agents and allows for the simultaneous consideration of microsystem improvements as well as macrosystem reforms. The open source release of our Java (Oracle, 2018) implementation of SiM-Care is currently in preparation.

3.3 Related Work

Decision support for health care planning is an area with increasing importance (Hamrock et al., 2013). To analyze health care systems, decision support tools have to deal with the detail complexity that is inherent to the health care sector: Patients schedule appointments based on their preferences and state of health, while PCPs offer appointments and treat patients. Micro-interactions thereby affect macro-level indicators as agents observe, learn and adapt, decide and act, and – as a group – determine the system's behavior (Gilbert, 2008). When a system's status depends on such micro-interactions, its behavior becomes difficult to predict and a "wide range of possible outcomes may arise from any policy change" (Fone et al., 2003). Simulation modeling can deal with this kind of complexity by "simulating the life histories of individuals and then estimating the population effect from the sum of the individual effects" (Fone et al., 2003). As such, simulation models represent a powerful tool to inform policy makers: They can provide valuable insights into the dependencies within health care systems and allow for the prediction of the outcome of a strategy change ahead of a potentially costly and risky real-world intervention (Fone et al., 2003; Hamrock et al., 2013).

Given these potentials, the use of computer simulation in health care delivery has significantly increased over the recent years (Zhong et al., 2016). The resulting body of literature is rich, as shown by several surveys of existing contributions. Examples include Fone et al. (2003) and Brailsford et al. (2009), who review the use of simulation modeling for health care in general. Other surveys are mostly focused on a particular simulation paradigm, e.g., system dynamics (SD) (Homer and Hirsch, 2006; Brailsford, 2008), discrete event simulations (DES) (Hamrock

et al., 2013; Jacobson et al., 2006), agent-based modeling (Barnes et al., 2013; Tracy et al., 2018), and hybrid simulations (Brailsford et al., 2019; Brailsford et al., 2010).

As background for the primary contribution of presenting a novel simulation system, we consider several examples of the computational study of primary care systems. The related references stem from a literature search featuring the keywords {*simulation, decision support, system dynamics, discrete event, agent based model*} + {*primary care, health care*}. Table 3.1 lists the resulting references and differentiates the simulation paradigm, the modeling objective, and information on stakeholder involvement and maintenance. Accordingly, we broadly partition the considered models into two groups: those studying microsystem improvements and those investigating macrosystem reforms. To allow for a direct comparison between these models and SiM-Care, we also include the latter in Table 3.1.

Simulation models aimed at studying microsystem improvements in primary care systems mostly include a detailed model of a single (specific) outpatient practice and focus on a predefined subset of potential improvements. Zhong et al. (2016) present a discrete event simulation for a pediatric clinic at the University of Wisconsin Health. Their model includes a very detailed representation of the sequential stages during a patient's visit. In a set of "what-if" scenarios, the authors investigate how the overall performance of the clinic is impacted by different scheduling templates, a change in the medical assistant to physician ratio, and the pairing of resident doctors with clinicians. Shi et al. (2014) develop a discrete event simulation model for a primary care clinic of the Department of Veteran Affairs. Within the model, the different patient flow routes for appointment patients, walk-ins, and nurse-only patients are distinguished. In a scenario analysis, the authors investigate how the clinic's performance is affected by six distinct factors that include walk-in and no-show rates as well as the double booking of appointments. Cayirli et al. (2006) use empirical data collected at a primary care clinic in New York to devise a discrete event simulation of a generic single-server primary care practice. The model distinguishes new and returning patients and accounts for walk-ins, no-shows, patient punctuality, and service time variations. In a simulation study, the authors evaluate 42 appointment systems that vary in the implemented sequencing- and appointment rules. A similar discrete event simulation of a generic single-server primary care practice is introduced by Schacht (2018). In the author's model, all arriving patients have a stochastic willingness to wait and always request an appointment. If the access time for this appointment exceeds a patient's willingness to wait, they become walk-ins. The arrival rate of patients depends on the session, day, and month to model seasonality. In a case study, the author evaluates a class of appointment systems that can account for seasonal variations in demand through reconfigurations. Further simulation models aimed at the study of microsystem improvements in primary care practices can be found in Wiesche et al. (2017), Oh et al. (2014), and Giachetti et al. (2005).

In contrast to SiM-Care, all of the models above include only one single primary care practice out of the many providers that make up a primary care system. Moreover, all of these models adopt a different approach to the representation of patients: While SiM-Care models a persistent patient population that is shared by all providers, the previous models perceive

Tab. 3.1.: Classification of related simulation models in primary care.

Reference	Method	Setting	Objective	Stakeholder Involvement	Maintenance
Cayirli et al., 2006	DES	single primary care clinic	eval. sequencing- and appointment rules	no information	no information, only management recommendations
Giachetti et al., 2005	DES	single outpatient clinic	testing a new scheduling approach	stakeholder involvement through action research	no information, implemented recommendations
Schacht, 2018	DES	single primary care clinic	eval. of appointment systems	no information	no information on maintenance, emphasize adaptability
Shi et al., 2014	DES	single primary care clinic	eval. effects of six factors on clinic's performance	management involvement in data collection	researchers provided only recommendations, no system
Wiesche et al., 2017	DES	single primary care clinic	eval. implications of capacity allocations and appointment scheduling	aimed to support stakeholders, no explicit involvement	no information on availability and maintenance
Zhong et al., 2016	DES	single pediatric clinic	eval. effects of scheduling templates, staff ratios, room assignments	analysis of exemplary clinic, no information on stakeholder involvement	no information
Homa et al., 2015	ABM	entire health care system	investigate paradox of primary care	cooperation between academics and patients, caregivers, and clinicians	model, software, and worksheets available for download and discussion
Matchar et al., 2016	SD	entire primary care sector	eval. effects of system-wide policy changes	group model building, development workshop	model handed over to Regional Health Systems
Comis et al., 2021	ABM + DES	entire primary care system	eval. quality of primary care systems as well as micro- and macrosystem changes	exchange with researchers, PCPs, health insurances, industry association, and administrative authorities	in preparation for open source release

patients as non-persistent, i.e., patients are generated as they arrive at the practice and cease to exist as soon as they are discharged. As a result, the previous models cannot account for the effects of individual microsystem improvements on the entire primary care system itself.

Simulation models that investigate macrosystem reforms of primary care systems mostly include an entire primary care system, however they are usually much more high level. To that end, Matchar et al. (2016) use the methodology of system dynamics to develop a simulation model to aid primary care planning in Singapore. The model captures the causal relationships between the stakeholders' aims and the provision of services in an analytical framework. The authors evaluate three policy changes that constitute in reducing the service gap, reducing the out-of-pocket costs, and increasing the number of physicians. While this and comparable system dynamics models do not consider the same level of micro-detail offered by agent-based simulations, they create fewer requirements with regard to computational effort and may provide a more concise model that is easier to communicate to stakeholders.

Homa et al. (2015) present an agent-based model to investigate the so-called paradox of primary care in which patients, PCPs, and specialists are represented as individual agents. Every patient has a health status that changes over time: The contraction of illnesses leads to a (temporary) decrease in the patients' health; the treatment of acute illnesses by PCPs and specialists as well as regular check-ups (performed exclusively by PCPs) lead to an increase in the patients' health. Tracking the evolution of the patients' average health status over time, the authors investigate how public health is affected by the interplay of different mechanisms in primary care. As such, their model has a different objective than SiM-Care: While the former investigates the external effects of treatments in primary care on the entire health care system, SiM-Care focuses on the processes within primary care systems. To that end, SiM-Care models the scheduling of appointments, the patients' actual practice visits that result in waiting times through the interaction of patients, and the physicians' treatments of patients with variable service times which are not part of the model by Homa et al. (2015).

To the best of our knowledge, there is no previous work on simulation models for the evaluation of primary care systems that allows for the simultaneous consideration of microsystem improvements and macrosystem reforms as in SiM-Care.

3.4 Outline and Use of Published Materials

Part I of this thesis is structured as follows. Chapter 4 introduces SiM-Care on the basis of the Overview, Design concepts, and Details (ODD) framework by Grimm et al. (2010). In Chapter 5, we subsequently present a case study based on empirical data to aid model validation and showcase how SiM-Care can be applied to support health care planning. Finally, Chapter 6 discusses the potential applications and entry requirements of SiM-Care and provides directions for future work.

All chapters of Part I are based on the publication Comis et al. (2021) and are thus joint work with my supervisors Christina Büsing and Catherine Cleophas.

4 A Simulation Model for Primary Care

Creating a simulation model means both formalizing what the model includes and deciding what to leave out (Tracy et al., 2018).. Therefore, this section first discusses the process of creating the model and the involvement of stakeholders before listing the resulting model assumptions and limitations. Subsequently, we delve into a more technical description of all modeled components and relationships based on the ODD framework described by Grimm et al. (2010).

SiM-Care is designed to meet the requirements of various stakeholders. *Researchers* access the model to evaluate outcomes from prescriptive planning approaches based on mathematical modeling. The modeling team regularly consulted with health care practitioners including *primary care physicians*, *health insurance representatives*, as well as *representatives from industry associations* and *administrative authorities*. Generally, we find that explaining the simulation model through the agent-based paradigm and presenting results from related studies allows for in-depth discussions, where the simulation provides a helpful tool for illustration.

At an early modeling stage, it became evident that the model would never be able to mirror all intricacies of a primary care system. Therefore, development focused on the idea of “modeling the problem, not the system", as recommended by Northridge and Metcalf (2016). Here, the primary problem is evaluating the macro-level effects from combining of health care supply in the form of a population of physicians versus a demand in the form of a population of patients. Thereby, we model the trade-offs between the objectives pursued by three stakeholder groups: patients, PCPs, and policy makers. SiM-Care assumes that PCPs strive to efficiently utilize their time, whereas patients strive for a quick response to their health concerns. Thereby, the model illustrates the trade-off between efficiency and patient-centered care. Policy objectives can range from minimizing the cost of health care to maximizing the degree of patient-centered care. Policy makers are not represented by agents within SiM-Care. Instead, policy decisions set relevant model parameters such as the number of physicians in the system and treatment standards. To model interactions on a micro-level, SiM-Care thus features two populations of agents: potential patients $\mathcal{P}$ and PCPs $\mathcal{G}$.

Every patient $\rho \in \mathcal{P}$ resides at a specific location, belongs to a certain age group and has an individual health status and treatment preferences; compare Figure 4.1. Patients develop acute illnesses that depend on their age and health status and require treatment. Additionally, patients may suffer from chronic illnesses, which need to be monitored by a physician. To receive medical attention, patients either schedule an appointment or visit a PCP’s practice

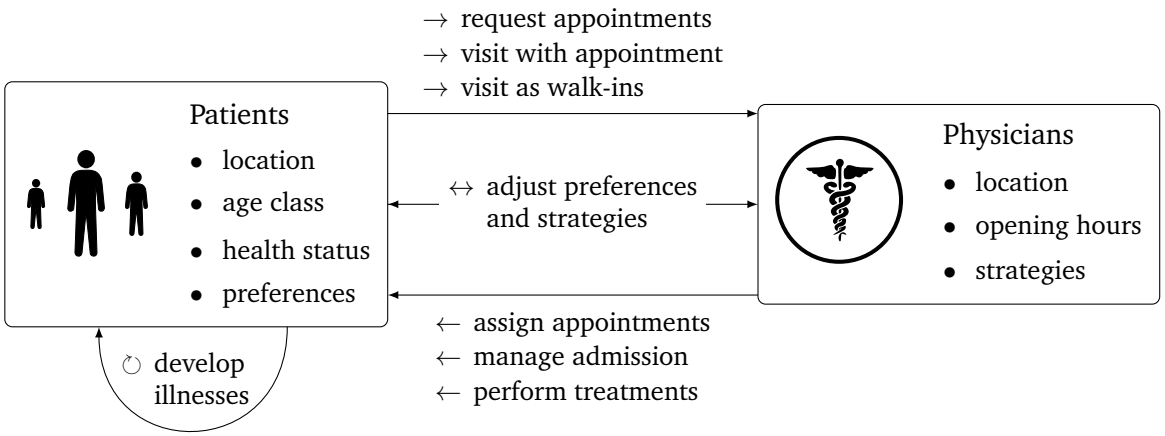


Fig. 4.1.: Concept of SiM-Care showing both types of agents with their main attributes as well as interactions between agents.

without prior notice. Patients' decisions depend on their individual preferences and health status. These factors determine the choice of physician, the type of the visit (walk-in/ appointment), and the time of the visit.

All PCPs $\phi \in \mathcal{G}$ practice at a certain location and have weekly recurring opening hours; see Figure 4.1. Moreover, every physician $\phi \in \mathcal{G}$ follows individual strategies that govern how they manage appointments, admit patients, and perform treatments. As patients and physicians interact, they influence each other and adjust their preferences and strategies.

Intrinsically, it is difficult – if not infeasible – to list all implicit assumptions underlying a model. Nevertheless, we list limitations that may restrict the application of SiM-Care in the following.

SiM-Care focuses on the adult population and neither models pediatric care nor gender differences. While we do differentiate patients by health status, age, and illnesses, we assume that all patients implement the same strategies when arranging appointments or becoming walk-ins. Furthermore, the model assumes that all patients attend their appointments, i.e., there are no no-shows patients.

As it stands, the model does not consider cross-effects between illnesses that may occur, e.g., when a chronic illness worsens the progression of an acute illness. As there is no model of direct patient interaction, SiM-Care does not include an explicit infection model, i.e., the probability of a patient developing an acute illness is independent of their interaction with other patients and physicians. While patients who suffer from illnesses seek treatment, the duration of an illness is not directly affected by treatments.

On the provider side, we do not model a relationship between primary care systems and specialists or hospitals. Physicians do not differentiate patients according to their insurance policy. The physicians' service times do not depend on the patients' number or types of illnesses and physicians do not offer home visits. We assume that PCPs are never late

or absent and the model includes neither seasonality nor holidays. Finally, we assume independence of surrounding municipalities, such that the modeled primary care system is a closed system.

To concisely highlight the interaction of the model's relevant components, we order and group the design questions given by the ODD framework as follows. First, Section 4.1 defines the temporal and geographical scales within SiM-Care. Second, Section 4.2 describes the representation of the relevant entities and state variables, including patients and physicians with their sensing, predicting, adapting, interacting, and learning actions. Third, Section 4.3 provides process overviews and describes matters of scheduling. Fourth, Section 4.4 explains how and where the model captures the uncertain nature of health care systems through stochastic parameters. Fifth, Section 4.5 briefly reviews the indicators that result from running the simulation and explains their emergent properties. Sixth, Section 4.6 discusses the initialization of a simulation experiment. Seventh, Section 4.7 documents the submodels that implement, e.g., the PCP's strategies to handle appointments. Finally, Section 4.8 describes our structural validation as well as our approach to verification taken when implementing the model.

4.1 Simulation Environment

SiM-Care's environment entails the geographical and temporal structure as well as policy effects. Within the model, locations $\ell \in L \coloneqq [-90, 90] \times [-180, 180]$ are represented using the geographic coordinates latitude and longitude indicating the north-south and east-west position, respectively.

The modeled time period is considered as a continuum structured by points in time and durations. For any time object (point in time or duration) $t = (\delta, \eta) \in \mathcal{T} \coloneqq \mathbb{N} \times [0, 1)$, attribute $\delta \in \mathbb{N}$ indicates the day and attribute $\eta \in [0, 1) =: \mathcal{H}$ specifies the time as an increment of day known as *decimal time*. That is, we use the same encoding for points in time and durations as context uniquely defines which of the former a time object refers to. For example, $(38, 0.55) \in \mathcal{T}$ corresponds to day 38 and $24 \cdot 60 \cdot 0.55 = 792$ minutes, i.e., 1:12 p.m. as a point in time or, analogously, to a duration of 38 days, 13 hours, and 12 minutes. To ease notation, we associate every point in time and duration $(\delta, \eta) \in \mathcal{T}$ with the non-negative value $\delta + \eta \in \mathbb{R}_{\geq 0}$ which yields a bijection between $\mathcal{T}$ and $\mathbb{R}_{\geq 0}$.

In addition to the continuous representation of time, we structure each day into a morning and an afternoon session as it is common practice in primary care (Klassen and Rohleder, 1996). Each *session* $\lambda = (\delta, \gamma) \in \Lambda \coloneqq \mathbb{N} \times \{0, 1\}$ is uniquely defined by a day $\delta \in \mathbb{N}$ and a binary indicator $\gamma \in \{0, 1\}$. Thereby, the binary indicator γ defines whether it is the morning

($\gamma = 0$) or the afternoon ($\gamma = 1$) session. Sessions recur on a weekly basis which yields an equivalence relation $\sim$ on the set of sessions Λ via

$$(\delta_1, \gamma_1) \sim (\delta_2, \gamma_2) :\Leftrightarrow \delta_1 \equiv \delta_2 \bmod 7 \ \wedge \ \gamma_1 = \gamma_2.$$

The resulting equivalence class for a session $\lambda \in \Lambda$ defined as $[\lambda] \coloneqq \{\lambda' \in \Lambda : \lambda' \sim \lambda\}$ contains all sessions sharing the same day of the week and time of the day, e.g., all Thursday afternoon sessions. Thus, we model and distinguish 14 sessions each week, i.e., Monday to Sunday with a respective morning and afternoon session that we associate with the set of all equivalence classes $\Lambda/\sim \coloneqq \{[\lambda] : \lambda \in \Lambda\}$. Particularly, this allows for a distinction between sessions on weekdays and weekends.

4.2 Entities and State Variables

Modeled as interacting agents, patients $\rho \in \mathcal{P}$ and PCPs $\phi \in \mathcal{G}$ are the active entities in the simulation. Their interaction is motivated by patients' suffering from illnesses and therefore seeking treatment with PCPs via appointments or walk-in visits. Both patients and PCPs are complex individuals featuring characteristics that represent entities themselves. Going from simple to more elaborated, we begin by describing the self-containing entities of SiM-Care and end with the description of the agents representing patients and physicians.

4.2.1 Objectives

When patients suffer from an acute illness, they want to be treated as soon as possible, ideally by their preferred physician. For the continuous treatment of chronic illnesses and the follow-up care of acute illnesses, patients prefer treatment by the same physician through appointments in regular intervals. Physicians, on the other hand, aim at efficiently utilizing their available time while minimizing overtime. Thus, patients' and physicians' objectives are in conflict as it is ineffective for physicians to fully comply with patient demands: To ensure that all short-notice appointment requests can be accommodated, PCPs would have to withhold too much treatment time. Providing follow-up appointments in strict intervals would prevent PCPs from reacting to demand fluctuations.

Policy makers, while not explicitly modeled, follow a multitude of conflicting objectives. On the one hand, they need to ensure a certain minimum standard in health care quality to guarantee patients are treated when necessary. On the other hand, they cannot afford to subsidize an excessive number of physicians. Thus, policy makers necessarily aim at a trade-off: A purely patient-based system that disregards efficiency is likely to turn out to be unaffordable, a health system optimized only for efficiency might lead to unacceptable waiting and access times. SiM-Care represents policy decisions through their resulting parameter values, e.g., the number of physicians and their distribution.

Tab. 4.1.: Summary of attributes and their units for illnesses $i \in \mathscr{I}$.

Attribute	Type	Unit
seriousness	$s_i \in [0,1]$	
illness family	$f_i \in \mathcal{F}$	
duration	$d_i \in \mathcal{T}$	[days]
willingness to wait	$\omega_i \in \mathcal{T}$	[days]
follow-up interval	$\nu_i \in \mathcal{T}$	[days]

4.2.2 Illnesses and Families of Illnesses

Illnesses are health concerns that cause discomfort to patients and require treatment. They belong to a certain illness family (e.g. cold or heartburn), have a certain seriousness (e.g. mild or severe), persist over a certain period of time, and require an initial treatment within an acceptable time frame as well as subsequent follow-up visits in regular time intervals. In SiM-Care, we formalize illnesses as tuples $i = (s_i, f_i, d_i, \omega_i, \nu_i) \in \mathscr{I}$ with attributes as shown in Table 4.1. Thereby, $s_i \in [0,1]$ defines the seriousness of the illness, $f_i \in \mathcal{F}$ defines the illness family of the illness, and $d_i \in \mathcal{T}$ defines the duration of the illness. The parameter $\omega_i \in \mathcal{T}$ defines the illness' willingness to wait, which is the patient's maximum accepted waiting time for the initial treatment. The parameter $\nu_i \in \mathcal{T}$ defines the illness' follow-up interval, which specifies the frequency of the required aftercare that follows the initial treatment. When we use this representation to model health concerns that are not strictly illnesses like the need for vaccination, the characteristics duration and follow-up interval may not apply. In such cases, setting parameter values $d_i = \emptyset$ and $\nu_i = \emptyset$ indicates that the respective characteristic is not applicable for $i \in \mathscr{I}$.

Families of illnesses serve as the classification system of illnesses within SiM-Care. While emerging illnesses vary in their manifestation, families of illnesses define the common constant traits of all illnesses belonging to the same illness family. In our model, the common constant traits of all illnesses $i \in \mathscr{I}$ with seriousness $s_i \in [0,1]$ belonging to illness family $f_i \in \mathcal{F}$ are the expected duration $D_{f_i}(s_i) \in \mathcal{T}$, the expected willingness to wait $W_{f_i}(s_i) \in \mathcal{T}$, and the follow-up interval $N_{f_i}(s_i) \in \mathcal{T}$. The expected duration $D_{f_i}(s_i)$ and expected willingness to wait $W_{f_i}(s_i)$ are exclusively used during the generation of new emerging illnesses and serve as the means for the distributions from which we sample each illness' stochastic duration d_i and stochastic willingness to wait ω_i. Thus for all emerged illnesses $i \in \mathscr{I}$, it generally holds that $d_i \neq D_{f_i}(s_i)$ and $\omega_i \neq W_{f_i}(s_i)$. Only the follow-up interval of emerged illnesses $i \in \mathscr{I}$ derives from the illness family in a deterministic way, i.e., $\nu_i = N_{f_i}(s_i)$.

In order to define the common traits of emerging illnesses, families of illnesses $f \in \mathcal{F}$ are formally specified by three functions. A linear function $D_f \colon [0,1] \to \mathcal{T}$ that defines the expected duration $D_f(s)$ in days for all emerging illnesses with seriousness $s \in [0,1]$ that derive from illness family $f \in \mathcal{F}$. Moreover, linear functions $W_f \colon [0,1] \to \mathcal{T}$ and $N_f \colon [0,1] \to \mathcal{T}$ that analogously define the expected willingness to wait in days and follow-up interval in days; see Table 4.2. As above, we indicate the inapplicability of the characteristics duration or follow-up interval to families of illnesses by setting $D_f = \emptyset$ and $N_f = \emptyset$, respectively.

Tab. 4.2.: Summary of attributes of families of illnesses $f \in \mathcal{F}$.

Attribute	Type
lin. function for exp. duration	$D_f \colon [0,1] \to \mathcal{T}$
lin. function for exp. willingness	$W_f \colon [0,1] \to \mathcal{T}$
lin. function for follow-up interval	$N_f : [0,1] \to \mathcal{T}$
chronic attribute	$\kappa_f \in \{0,1\}$

To illustrate the concept of illnesses and families of illnesses, consider the illness family "common cold" with expected illness duration defined by $D_f(s) = 10\,s + 3$, expected willingness to wait defined by $W_f(s) = -3\,s + 3$, and follow-up interval defined by $N_f(s) = -2\,s + 7$. When a patient develops a mild ($s_i = 0.2$) "common cold", the illness family "common cold" defines the expected duration, expected willingness to wait, and follow-up interval of the mild cold as $D_f(s_i) = 5$ days, $W_f(s_i) = 2.4$ days, and $N_f(s_i) = 6.6$ days. The actual duration and willingness to wait of the developed mild "common cold" are stochastic and vary around their expected counterparts, e.g., $d_i = 5.5$ days and $\omega_i = 2.7$ days. The illness' follow-up interval is deterministic and derives from the illness family via $\nu_i = N_{f_i}(s_i) = 6.6$ days. The particular mild cold in this example will thus not require a follow-up visit as its duration is shorter than the follow-up interval, i.e., $d_i < \nu_i$.

To model chronic health concerns such as diabetes that persist over an extended period of time, a chronic attribute $\kappa_f \in \{0,1\}$ identifies families of chronic illnesses. Thereby, κ_f partitions $\mathcal{F}$ into the set of acute families of illnesses $\mathcal{F}^{\text{act}} \coloneqq \{f \in \mathcal{F} : \kappa_f = 0\}$ and the set of chronic families of illnesses $\mathcal{F}^{\text{chro}} \coloneqq \{f \in \mathcal{F} : \kappa_f = 1\}$. This directly induces a partition of the set of illnesses $\mathscr{I}$ into the set of acute $\mathscr{I}^{\text{act}}$ and the set of chronic illnesses $\mathscr{I}^{\text{chro}}$.

Acute illnesses $i \in \mathscr{I}^{\text{act}}$ develop and subside over time and patients $\rho \in \mathcal{P}$ can simultaneously suffer from an arbitrary number of acute illnesses $\mathcal{I}_\rho^{\text{act}} \subseteq \mathscr{I}^{\text{act}}$. Chronic illnesses $\varsigma \in \mathscr{I}^{\text{chro}}$ are conceived as static by the model – they neither develop nor heal in the modeled time period. Instead, each patient $\rho \in \mathcal{P}$ either suffers from exactly one chronic illness $\varsigma_\rho \in \mathscr{I}^{\text{chro}}$ throughout the modeled time period, i.e., $\mathcal{I}_\rho^{\text{chro}} = \{\varsigma_\rho\} \subseteq \mathscr{I}^{\text{chro}}$, or no chronic illness at all, i.e., $\mathcal{I}_\rho^{\text{chro}} = \emptyset$. To distinguish patients suffering from a chronic illness from those who do not, we refer to the former as *chronic patients*.

4.2.3 Appointments

Appointments specify the point in time at which the treatment of a specific patient is scheduled to take place. To that end, appointments $b \in \mathcal{B}$ are defined by the time of the appointment $t_b \in \mathcal{T}$, the attending physician $\phi_b \in \mathcal{G}$, and the patient $\rho_b \in \mathcal{P}$ receiving treatment. At any point in time, non-chronic patients can have at most one scheduled appointment $b^{\text{act}} \in \mathcal{B}$, called the *acute* appointment. Acute appointments are intended for the initial treatment of acute illnesses, the follow-up treatment of acute illnesses, or both. Chronic patients, may have a *regular* appointment $b^{\text{reg}} \in \mathcal{B}$ to treat their chronic illness in addition to the acute appointment to treat their acute illnesses. While chronic illnesses are only treated during

Tab. 4.3.: Summary of attributes of age classes $a \in \mathcal{A}$.

Attribute	Type
lin. function for exp. annual acute illnesses	$I_a \colon [0,1] \to \mathbb{R}_{\geq 0}$
deviation from exp. illness duration	$\Delta_a^d > 0$
deviation from exp. willingness to wait	$\Delta_a^\omega \geq 0$
probability to cancel appointments	$p_a \in [0,1]$

regular appointments, acute illnesses are treated during any appointment. Thus, all of a patients' acute illnesses $\mathcal{I}^{\text{act}}$ are treated during every appointment.

4.2.4 Age Classes

Age classes group the modeled set of patients and serve the purpose of defining the common characteristics of patients within the respective classes. For patients of age class $a \in \mathcal{A}$, these characteristics are the deviation from the expected illness duration $\Delta_a^d > 0$, the deviation from the expected willingness to wait $\Delta_a^\omega \geq 0$, the probability to cancel an appointment after full recovery $p_a \in [0,1]$, and the expected number of annual acute illnesses defined through the linear function $I_a \colon [0,1] \to \mathbb{R}_{\geq 0}$; see Table 4.3. The deviation from the expected illness duration Δ_a^d is a multiplicative factor, that determines whether the expected illness duration $D_{f_i}(s_i) \in \mathcal{T}$ extends ($\Delta_a^d > 1$) or shortens ($\Delta_a^d < 1$) for patients of age class $a \in \mathcal{A}$. Analogously, the deviation from the expected willingness to wait Δ_a^ω, determines how the expected willingness to wait $W_{f_i}(s_i) \in \mathcal{T}$ of an illness changes for patients of age class $a \in \mathcal{A}$. The linear function $I_a \colon [0,1] \to \mathbb{R}_{\geq 0}$ defines the expected number of annual acute illnesses $I_a(c) \in \mathbb{R}_{\geq 0}$ for patients in age class $a \in \mathcal{A}$ which depends on the patient's individual health condition $c \in [0,1]$ that can range from perfectly healthy ($c = 0$) to extremely delicate ($c = 1$).

4.2.5 Age Class-Illness Distribution

The age class-illness distribution $\pi^{\text{act}} \colon \mathcal{A} \times \mathcal{F}^{\text{act}} \to [0,1]$ builds the connection between the set of age classes $\mathcal{A}$ and the set of acute families of illnesses $\mathcal{F}^{\text{act}}$. To that end, π^{act} defines the expected distribution of acute illness families for each age class, i.e., among all developed acute illnesses by patients of age class $a \in \mathcal{A}$, a fraction $\pi^{\text{act}}(a, f_i) \in [0,1]$ is expected to belong to illness family $f_i \in \mathcal{F}^{\text{act}}$. As a result, π^{act} defines a discrete probability distribution on the set of acute families of illnesses $\mathcal{F}^{\text{act}}$ for fixed age class $a \in \mathcal{A}$, i.e., $\sum_{f_i \in \mathcal{F}^{\text{act}}} \pi^{\text{act}}(a, f_i) = 1$.

4.2.6 Patients

Patients are the driving force of the simulation, as their health concerns trigger the events that underly most of the simulation's processes. All non-chronic patients $\rho \in \mathcal{P}$ are characterized by their geographical location $\ell_\rho \in L$, health condition $c_\rho \in [0,1]$, acute illnesses $\mathcal{I}_\rho^{\text{act}} \subseteq \mathscr{I}^{\text{act}}$,

age class $a_\rho \in \mathcal{A}$, acute appointment $b_\rho^{\text{act}} \in \mathcal{B}$, and preferences. While the location, health condition, and age class of each patient remain constant throughout a simulation experiment, a patient's acute illnesses, acute appointment and preferences are variable and change over time. Chronic patients possess all the characteristics of non-chronic patients, but are additionally identified by a constant chronic illness $\mathcal{I}_\rho^{\text{chro}} = \{\varsigma_\rho\} \subseteq \mathscr{I}^{\text{chro}}$ and a variable regular appointment $b_\rho^{\text{reg}} \in \mathcal{B}$. We denote all of a patient's illnesses by $\mathcal{I}_\rho \coloneqq \mathcal{I}_\rho^{\text{act}} \cup \mathcal{I}_\rho^{\text{chro}}$.

Patients' preferences determine when, where and how they pursue treatment. Specifically, each patient $\rho \in \mathcal{P}$ considers a set of PCPs $\mathcal{G}_\rho^{\text{con}} \subseteq \mathcal{G}$ and never seeks treatment with PCPs outside the consideration set. Since continuity in the treatment of chronic illnesses is particularly important, chronic patients select a distinguished family physician $\phi_\rho^{\text{fam}} \in \mathcal{G}_\rho^{\text{con}}$ with whom all regular appointments $b_\rho^{\text{reg}} \in \mathcal{B}$ are exclusively arranged. While every patients' consideration set $\mathcal{G}_\rho^{\text{con}}$ remains constant throughout the modeled time period, patients reevaluate and vary their family physician. Naturally, patients have personal schedules and cannot attend all weekly sessions. Thus, the model assumes that each patient has a constant set of weekly-recurring session availabilities given by $\alpha_\rho \colon \Lambda/\!\sim \to \{0, 1\}$, where 0 encodes unavailability. Finally, every patient $\rho \in \mathcal{P}$ maintains an individual appointment rating $r_\rho^{\text{app}}(\phi) \geq 0$ as well as a session-specific walk-in rating $r_\rho^{\text{walk}}(\phi, [\lambda]) \geq 0$ for every weekly session $[\lambda] \in \Lambda/\!\sim$ and every considered physician $\phi \in \mathcal{G}_\rho^{\text{con}}$.

Ratings are the means by which patients express their satisfaction with a physician's services. Whenever a patient seeks consultation, the choice of physician is determined by the patient's current ratings. To that end, ratings incorporate the patients' sense of geographic distance, matching of opening hours with availabilities, and previous positive and negative experiences. As patients adjust their ratings over time, they adjust their choice of PCP. If a physician is unable to meet an appointment request, causes excessive waiting times, or rejects patients due to capacity overruns, patients reduce their rating. Positive experiences such as successful appointment arrangements or short waiting times increase ratings. In other terms, through their sensing of the quality of treatment and the adaptation of their ratings, patients learn about the quality of PCPs throughout the simulation cycle.

When patients begin to suffer from a new illness, they always seek treatment. To that end, patients $\rho \in \mathcal{P}$ first request an appointment from the set of considered PCPs $\mathcal{G}_\rho^{\text{con}}$. Appointment requests are one of the ways in which patients and PCPs interact. Patients attempt up to two appointment requests in order of the appointment rating $r_\rho^{\text{app}}(\phi) \geq 0$ they assign to the considered physicians $\phi \in \mathcal{G}_\rho^{\text{con}}$. If both requested PCPs fail to offer a feasible appointment within the patient's willingness to wait, patients resort to their second way of interacting with physicians: They forgo an appointment and visit a PCP as a walk-in. The selection of the PCP for the walk-in visit is based on the corresponding walk-in rating $r_\rho^{\text{walk}}(\phi, [\lambda])$ of the targeted session $\lambda \in \Lambda$.

Upon arrival, patients may be rejected by physicians due to, e.g., capacity overloads. Following a rejection, patients update their rating of the rejecting PCP and attempt a new visit as walk-in at the then-highest-rated PCP. Rejected patients are flagged as emergencies ($\varepsilon_\rho = 1$) for as

Tab. 4.4.: Summary of attributes of (chronic) patients $\rho \in \mathcal{P}$.

Attribute	Domain	Type
location	$\ell_\rho \in L$	constant
health condition	$c_\rho \in [0, 1]$	constant
age class	$a_\rho \in \mathcal{A}$	constant
acute illnesses	$\mathcal{I}_\rho^{\text{act}} \subseteq \mathscr{I}^{\text{act}}$	variable
emergency flag	$\varepsilon_\rho \in \{0, 1\}$	variable
acute appointment	$b_\rho^{\text{act}} \in \mathcal{B}$	variable
considered PCPs	$\mathcal{G}_\rho^{\text{con}} \subseteq \mathcal{G}$	constant
availabilities	$\alpha_\rho \colon \Lambda/\sim \to \{0, 1\}$	constant
appointment ratings	$r_\rho^{\text{app}}(\phi) \geq 0,\ \forall \phi \in \mathcal{G}_\rho^{\text{con}}$	variable
walk-in ratings	$r_\rho^{\text{walk}}(\phi, [\lambda]) \geq 0,\ \forall \phi \in \mathcal{G}_\rho^{\text{con}},\ \forall [\lambda] \in \Lambda/\sim$	variable
chronic illness	$\mathcal{I}_\rho^{\text{chro}} = \{\varsigma_\rho\} \subseteq \mathscr{I}^{\text{chro}}$	constant
regular appointment	$b_\rho^{\text{reg}} \in \mathcal{B}$	variable
family physician	$\phi_\rho^{\text{fam}} \in \mathcal{G}_\rho^{\text{con}}$	variable

long as they unsuccessfully continue to seek treatment. In our model, this emergency state does not enforce a particular PCP behavior. Instead, PCPs may include the emergency state in their decision making.

Until an illness $i \in \mathcal{I}_\rho^{\text{act}}$ subsides, patients continuously try to arrange follow-up appointments to the initial treatment with the attending physician in the follow-up interval $\nu_i \in \mathcal{T}$. Analogously, chronic patients $\rho \in \mathcal{P}$ continuously try to arrange regular appointments with their family physician $\phi_\rho^{\text{fam}} \in \mathcal{G}_\rho^{\text{con}}$ in the follow-up interval $\nu_{\varsigma_\rho} \in \mathcal{T}$ of their unique chronic illness $\varsigma_\rho \in \mathcal{I}_\rho^{\text{chro}}$. Only if the arrangement of a follow-up or regular appointment fails and the aftercare is endangered, do patients seek follow-up treatment as walk-ins. As a result, a patient's chronic illness $\varsigma_\rho \in \mathcal{I}_\rho^{\text{chro}}$ can be treated by a physician other than the family physician $\phi_\rho^{\text{fam}} \in \mathcal{G}_\rho^{\text{con}}$, but only through a walk-in visit triggered by the unavailability of a regular appointment.

In SiM-Care, patients do not directly interact with other patients. However, an indirect form interaction emerges as patients compete with each other for timely treatment by their preferred PCP.

The attributes shared by all patients as well as the attributes specific to chronic patients are summarized in Table 4.4. In the following, we will regularly omit the indices of the patients' attributes to ease notation.

4.2.7 Primary Care Physicians

PCPs operate practices featuring an uncapacitated waiting room to offer medical services to patients in need. The model characterizes physicians $\phi \in \mathcal{G}$ by their geographic location $\ell_\phi \in L$, opening hours, as well as an individual set of strategies to schedule appointments, manage patient admissions, and organize treatments.

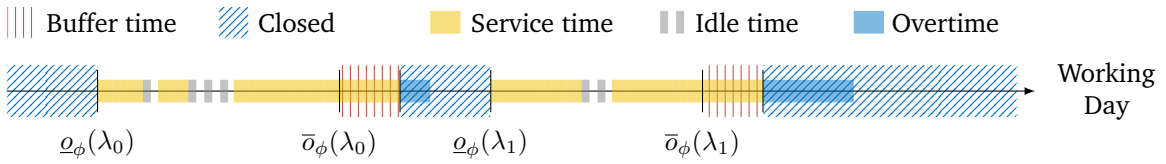


Fig. 4.2.: Schematic representation of the morning (λ_0) and afternoon (λ_1) session of PCP $\phi \in \mathcal{G}$ visualizing service-, idle- and overtime.

SiM-Care assumes that all physicians $\phi \in \mathcal{G}$ operate in clinical sessions. Opening hours for these sessions are weekly recurring and therefore defined for the sessions of the week via $o_\phi \colon \Lambda/\sim \to \mathcal{H} \times \mathcal{H}$ where $\mathcal{H}$ denotes the set of decimal times defined in Section 4.1. Opening hours specify for each session $\lambda \in \Lambda$ the time window $o_\phi([\lambda]) \coloneqq [\underline{o}_\phi([\lambda]), \overline{o}_\phi([\lambda])]$ during which physician $\phi \in \mathcal{G}$ generally admits patients for treatment. The beginning of session $\lambda = (\delta, \gamma) \in \Lambda$ is defined as $\underline{o}_\phi(\lambda) \coloneqq (\delta, \underline{o}_\phi([\lambda])) \in \mathcal{T}$, the session's end as $\overline{o}_\phi(\lambda) \coloneqq (\delta, \overline{o}_\phi([\lambda])) \in \mathcal{T}$. To encode that PCP $\phi \in \mathcal{G}$ is closed for a weekly session $[\lambda] \in \Lambda/\sim$, we set $o_\phi([\lambda]) = \emptyset$. Physicians utilize the first hour after the end of each session as time buffer to compensate for possible delays and walk-ins. Buffers are considered anticipated working time so that only service time that extends beyond the buffer constitutes overtime. Figure 4.2 provides a schematic visualization of a PCP's working day.

PCPs implement a set of *strategies* to schedule appointments, decide on patient admissions, and organize the treatment of patients. These strategies govern the physicians' interactions with patients and incorporate all of their sensing, predicting, adapting, and learning.

The PCP's *appointment scheduling strategy* $S_\phi \in \mathcal{S}^{\text{app}}$ defines how consultation time is allocated to appointment slots and how the resulting slots are assigned to requesting patients. The feasible set of appointment scheduling strategies $\mathcal{S}^{\text{app}}$ is defined via the interface shown in Figure 4.3. That is, every appointment scheduling strategy $S \in \mathcal{S}^{\text{app}}$ has to provide the functionality to answer appointment requests with an appointment suggestion (that can be empty). Thereby, every appointment request specifies the requesting patient, earliest possible appointment time, willingness to wait, whether the request is for a regular appointment, and whether the patient's availabilities have to be respected. Furthermore, every appointment scheduling strategy $S \in \mathcal{S}^{\text{app}}$ has to provide the functionality to schedule previously offered appointments as well as the functionality to cancel previously scheduled appointments. Finally, every appointment scheduling strategy $S \in \mathcal{S}^{\text{app}}$ has to be able to compute the number of upcoming appointments within a session that are scheduled to take place after a specified point in time.

The PCP's *treatment strategy* $S_\phi \in \mathcal{S}^{\text{tmt}}$ defines the order of treatment among patients from the waiting room. Physicians sense their patients' waiting times as input for their strategy. To account for the observation that physicians consciously or unconsciously adjust service times depending on demand (Gupta and Denton, 2008), treatment policies define when and how physicians adjust their consultation speed and thereby service times. The feasible set of treatment strategies $\mathcal{S}^{\text{tmt}}$ is defined via the interface shown in Figure 4.3. That is, every

```
1  public interface IAppointmentSchedulingStrategy {
2    public Optional<Appointment> findAppointment(AppointmentRequest
         rq);
3    public void scheduleAppointment(Appointment b);
4    public void cancelAppointment(Appointment b);
5    public int upcomingAppointmentAfter(Time t);
6  }
7
8  public interface ITreatmentStrategy {
9    public void handleArrival(ArrivalEvent ae);
10   public int[] waitingPatients();
11   public void sessionStarted();
12   public Optional<ArrivalEvent> getNextPatient();
13   public float getConsultationSpeed();
14 }
15
16 public interface IAdmissionStrategy {
17   public boolean acceptPatient(ArrivalEvent ae,
         IAppointmentSchedulingStrategy as , ITreatmentStrategy ts);
18   public void adaptPolicy(Session session, ITreatmentStrategy ts);
19 }
```

Fig. 4.3.: Interfaces implemented by PCPs' strategies.

Tab. 4.5.: Summary of attributes of PCPs $\phi \in \mathcal{G}$.

Attribute	Type
location	$\ell_\phi \in L$
opening hours	$o_\phi \colon \Lambda/\sim \to \mathcal{H} \times \mathcal{H}$
appt. scheduling strategy	$S_\phi \in \mathcal{S}^{\text{app}}$
admission strategy	$S_\phi \in \mathcal{S}^{\text{adm}}$
treatment strategy	$S_\phi \in \mathcal{S}^{\text{tmt}}$

treatment strategy $S \in \mathcal{S}^{\text{tmt}}$ has to keep track of admitted patients, count the number of waiting patients with and without appointment, and define how the treatment strategy is affected by the beginning of a session. Moreover, every treatment strategy $S \in \mathcal{S}^{\text{tmt}}$ has to determine the next patient to be treated (that might not exist) as well as the PCP's current consultation speed which is thoroughly discussed in Section 4.7.4.

The PCP's *admission strategy* $S_\phi \in \mathcal{S}^{\text{adm}}$ determines whether a physician admits or rejects an arriving patient based on the current workload. Admitted patients await their treatment in the physician's waiting room. In SiM-Care, PCPs are required to treat all admitted patients. Thus, physicians underestimating their workload due to faulty predictions might have to work overtime as they accept too many patients. On the other hand, physicians that overestimate their workload reject too many patients and fail to fully utilize their available time. At the end of every session's buffer, physicians learn by reevaluating their predictions and adapting their admission policy. The feasible set of admission strategies $\mathcal{S}^{\text{adm}}$ is defined via the interface shown in Figure 4.3. That is, every admission strategy $S \in \mathcal{S}^{\text{adm}}$ has to be able to decide whether an arriving patient is admitted or not given the PCP's treatment and appointment scheduling strategy. Moreover, every admission strategy $S \in \mathcal{S}^{\text{adm}}$ has to define the adaptive traits that are performed at the end of every session's buffer and depend on the PCP's treatment strategy.

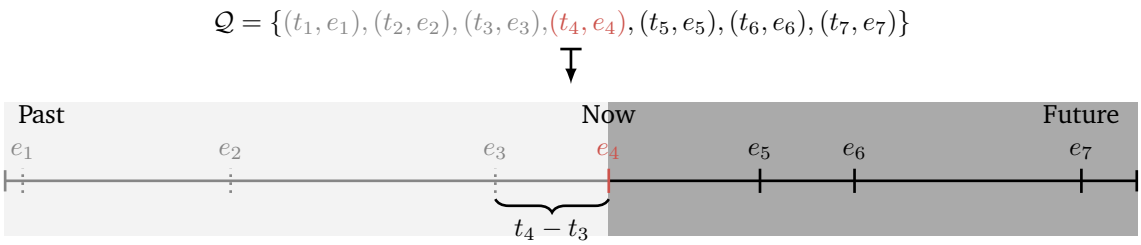


Fig. 4.4.: Progression of time induced by the processing of event queue $\mathcal{Q}$ via the discrete event paradigm.

Physicians do not directly interact with other physicians. However, an indirect form of interaction emerges as PCPs compete for the patients' favor while striving for optimal utilization.

The attributes of PCP's are summarized in Table 4.5. As for patients, we generally omit the indices for the physicians' attributes to ease notation.

4.3 Process Overview and Scheduling

Within SiM-Care, the progression of time is modeled via the discrete event paradigm. That is, time is a continuum which is traversed between discrete events at which the system state is updated. The model stores events of the form (t, e) in a sequential queue $\mathcal{Q}$ where $t \in \mathcal{T}$ is the point in time an event of type $e \in \mathcal{E}$ occurs.

Events in $\mathcal{Q}$ happen chronologically, i.e., $\mathcal{Q} = \{(t_1, e_1), \ldots, (t_n, e_n)\}$ with $t_i \leq t_{i+1}$ for $1 \leq i \leq n-1$. As soon as an event $(t_i, e_i) \in \mathcal{Q}$ occurs, the simulation advances from time t_{i-1} to time t_i, compare Figure 4.4. The simulation terminates at a specified point in time $T \in \mathcal{T}$, i.e., when the first element $(t_i, e_i) \in \mathcal{Q}$ with $t_i \geq T$ occurs.

Any event $(t, e) \in \mathcal{Q}$ can generate new events or delete existing ones. To be introduced or affected by event $(t, e) \in \mathcal{Q}$, events $(t', e') \in \mathcal{Q}$ must happen after time t, i.e., we require $t' > t$, so that time progresses in a consistent fashion.

By construction, $\mathcal{Q}$ never runs empty. Every simulation run follows the structure depicted in Figure 4.5, chronologically processing events in $\mathcal{Q}$ until time $T \in \mathcal{T}$ is reached. In this, the specific process depends on the event type $e \in \mathcal{E}$. We now detail these event types.

Arrival events are indicated by $e^{\text{arv}}(\phi, \rho)$. As illustrated in Figure 4.6(a), they mark the event of patient ρ arriving at the practice of physician ϕ, either for an appointment or as a walk-in. The physician's decision to admit or reject arriving patients depends on admission strategy of ϕ. Every admitted patient is guaranteed to receive treatment and enters the physician's waiting room. If the physician is currently idle, this triggers the physician's treatment strategy and treatment commences.

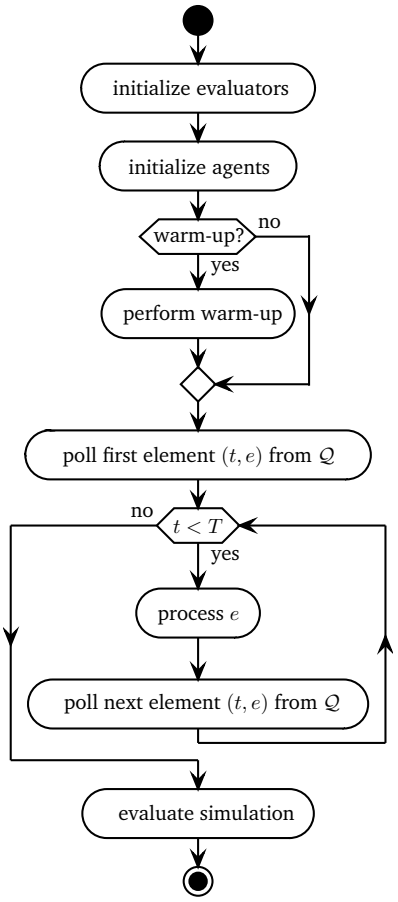


Fig. 4.5.: Structure of simulation run with time horizon T.

Follow-up events are indicated by $e^{\text{fol}}(\phi, \rho, i)$. Some families of illnesses $f_i \in \mathcal{F}$ cannot be treated via a single visit. Instead, the related illnesses $i \in \mathcal{I}$ require follow-up treatments at a frequency defined by the parameter $\nu_i \neq \emptyset$. Ensuring continuous follow-up treatments, patients always try to arrange a follow-up appointment immediately after the treatment of illnesses requiring follow-up consultation. To account for the fact that no feasible follow-up appointment might be available, SiM-Care generates a follow-up event $e^{\text{fol}}(\phi, \rho, i)$ at time $t^{\text{treat}} + \nu_i$ every time illness $i \in \mathcal{I}$ with $\nu_i \neq \emptyset$ suffered by patient $\rho \in \mathcal{P}$ is treated by physician $\phi \in \mathcal{G}$ at time $t^{\text{treat}} \in \mathcal{T}$. Follow-up events serve as the patient's reminder to actively re-pursue follow-up consultation for illness i after the duration of the follow-up interval. Triggered by a follow-up event $e^{\text{fol}}(\phi, \rho, i)$, patient ρ reattempts to arrange a follow-up appointment with physician ϕ. Should ϕ once again be unable to provide a suitable appointment, ρ seeks follow-up consultation as a walk-in; compare Figure 4.6(b). Every follow-up treatment of an illness $i \in \mathcal{I}$ invalidates all associated existing follow-up events, as the follow-up interval is reset. Therefore, SiM-Care deletes all existing follow-up events $e^{\text{fol}}(\phi, \rho, i) \in \mathcal{Q}$ associated with illnesses $i \in \mathcal{I}$ that were treated during a visit before the new follow-up events are generated. As a result, follow-up events only trigger if an illness has not been treated for the duration of its follow-up interval $\nu_i \in \mathcal{T}$.

Release events are indicated by $e^{\text{rel}}(\phi, \rho)$. As illustrated in Figure 4.6(c), release events mark the event of physician ϕ releasing patient ρ after a treatment is performed. Whenever a new treatment begins, the sampled service time determines the time of the subsequent release event $e^{\text{rel}}(\phi, \rho)$. All treated illnesses $i \in \mathcal{I}^{\text{act}}$ without duration $(d_i = \emptyset)$ are cured through a one-time treatment and thus removed from $\mathcal{I}^{\text{act}}$. Subsequently, all existing follow-up events corresponding to treated illnesses are deleted and new follow-up events are generated in the previously described manner. The successful treatment revokes existing emergency flags, i.e., we set $\varepsilon = 0$. If the patient's chronic illness $\varsigma \in \mathcal{I}$ was treated, the next recurrent regular appointment $b^{\text{reg}} \in \mathcal{B}$ with physician ϕ^{fam} is requested at time $t^{\text{treat}} + \nu_\varsigma$. Then, patients request an acute appointment $b^{\text{act}} \in \mathcal{B}$ with physician ϕ for the follow-up treatment of the persisting acute illness $i^* = \text{argmin}_{i \in \mathcal{I}^{\text{act}}: \nu_i \neq \emptyset} \nu_i$ with smallest follow-up interval. The requested appointments ensure the follow-up treatment of all illnesses suffered by patient ρ and will preempt the previously generated follow-up events. Finally, physicians implement

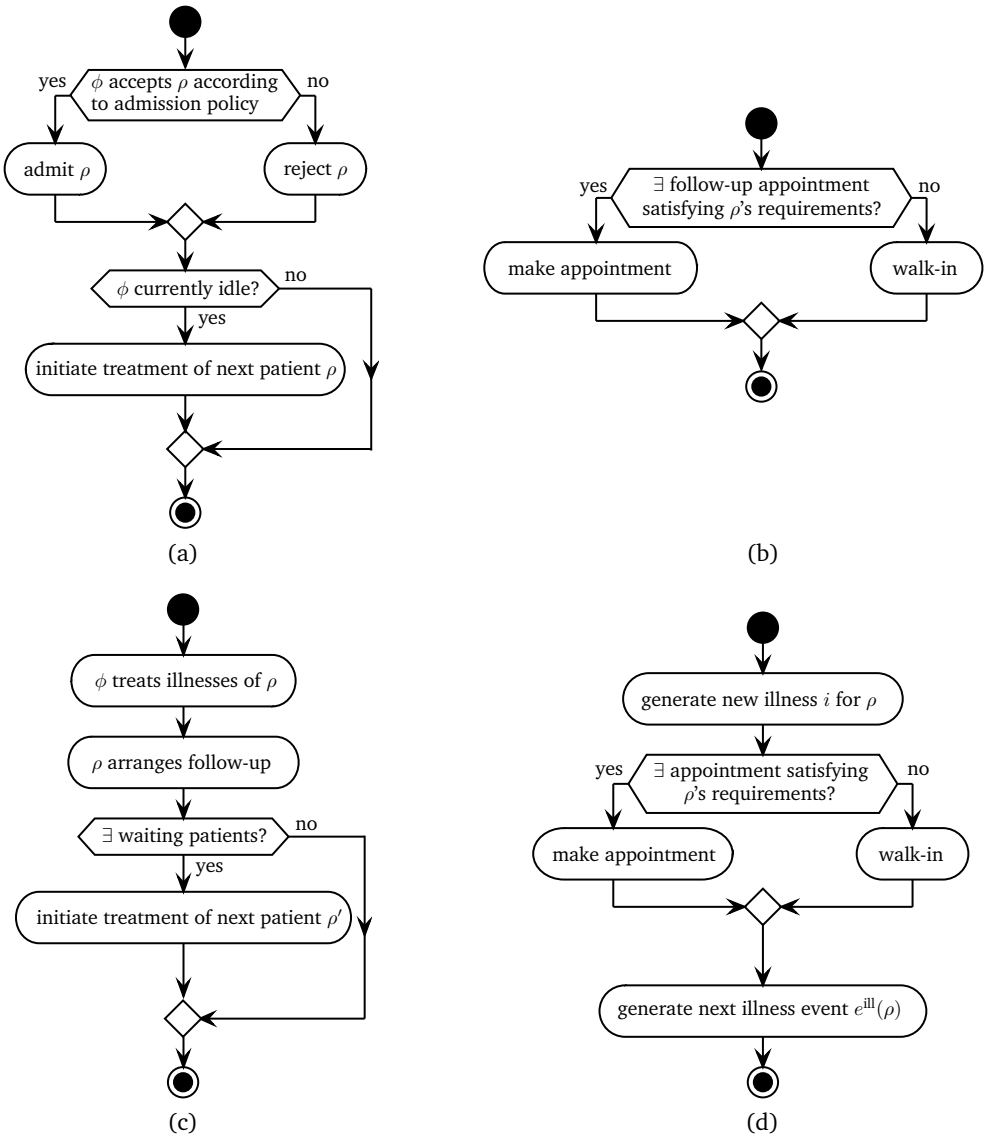


Fig. 4.6.: Processing of (a) arrival events $e^{\text{arv}}(\phi, \rho)$, (b) follow-up events $e^{\text{fol}}(\phi, \rho, i)$, (c) release events $e^{\text{rel}}(\phi, \rho)$, and (d) illness events $e^{\text{ill}}(\rho)$; $\phi \in \mathcal{G}$, $\rho \in \mathcal{P}$, and $i \in \mathcal{I}$.

their treatment strategy to select the next patient from the waiting room if the latter is non-empty. Otherwise, physicians remain idle until the next arrival event triggers the treatment strategy. As a result of this behavior, physicians are never intentionally idle.

Illness events are indicated by $e^{\text{ill}}(\rho)$. As illustrated in Figure 4.6(d), they describe that patient ρ starts to suffer from a new acute illness. This means that the model generates a new acute illness $i \in \mathscr{I}^{\text{act}}$ with stochastic qualities that depend on the patient's age and health condition and adds it to the patient's set of illnesses $\mathcal{I}$. To treat emerged illnesses, patients request an appointment from their preferred physicians or, in case this does not succeed, directly visit the preferred physician as a walk-in. As a result, each illness event generates a corresponding arrival event $e^{\text{arv}}(\phi, \rho)$ and adds it to the queue $\mathcal{Q}$. Finally, each illness event

generates a future illness event $e^{\text{ill}}(\rho)$ for patient ρ and adds it to the queue $\mathcal{Q}$ to mark the next point in time patient ρ develops an acute illness.

Recovery events are indicated by $e^{\text{rec}}(\rho, i)$. They mark the event of patient ρ recovering from acute illness $i \in \mathcal{I}^{\text{act}}$. Whenever the model generates a new acute illness $i \in \mathscr{I}^{\text{act}}$ with $d_i \neq \emptyset$, it also generates a corresponding recovery event $e^{\text{rec}}(\rho, i)$ at time $t^{\text{ill}} + d_i$, where $t^{\text{ill}} \in \mathcal{T}$ is the point in time illness i is developed. Illnesses without duration ($d_i = \emptyset$) do not require a recovery event as they are immediately cured through their initial treatment. A recovery event removes illness i from $\mathcal{I}$ and deletes any associated follow-up event $e^{\text{fol}}(\phi, \rho, i) \in \mathcal{Q}$. If patient ρ does not suffer from acute illnesses following the removal of illness i, i.e., $\mathcal{I}^{\text{act}} = \emptyset$, the model revokes existing emergency flags by setting $\varepsilon = 0$ and assumes that ρ may cancel scheduled acute appointments. Such cancellations occur with the patient's age-class specific probability $p_a \in [0, 1]$ and consequently delete the associated arrival event $e^{\text{arv}}(\phi, \rho)$. As a result, some patients keep their existing acute appointment for a final debriefing. Should patient ρ be currently seeking walk-in treatment due to persisting chronic illness $\varsigma \in \mathcal{I}$, this effort is continued. Otherwise, the current walk-in attempt is canceled and the associated arrival event $e^{\text{arv}}(\phi, \rho)$ is deleted.

Open- and close events are indicated by $e^{\text{opn}}(\phi)$ and $e^{\text{clo}}(\phi)$, respectively. They mark the beginning and ending (including buffer) of a session $\lambda \in \Lambda$ operated by physician ϕ. They ensure that treatment strategies become aware of a session's beginning, e.g., to allow for strategies that do not treat early-arriving patients before $\underline{o}(\lambda)$, and that overtime is incurred for all treatments performed beyond the anticipated buffer time of λ.

4.4 Modeling Variability

SiM-Care relies on stochastic values to both approximate real-world variability and control the frequency of events. This applies to aspects of illnesses as well as to patient arrivals, appointment cancellations and service times. In consequence, every simulation experiment includes multiple stochastic repetitions of the modeled time period, termed *simulation runs*. When examining simulation output, we account for the resulting variability through confidence intervals.

Table 4.6 lists all aspects of the model that are probabilistic. In the following, we detail the parameterization of the distributions underlying the random values.

Frequency of Acute Illnesses. The occurrence of acute illnesses in SiM-Care is modeled via a Poission process. Patients develop acute illnesses at a frequency that depends on their age and health condition. For patients $\rho \in \mathcal{P}$ of age class $a \in \mathcal{A}$ with health condition $c \in [0, 1]$, the expected number of acute illnesses per year is given by the parameter $I_a(c)$.

Tab. 4.6.: Probabilistic model aspects.

Aspect	Distribution
frequency of acute illnesses	exponential dist.
type of acute illnesses	age class-illness dist.
seriousness of acute illnesses	triangular dist.
duration of acute illnesses	log-normal dist.
patients' willingness to wait	Weibull dist.
patient punctuality	normal dist.
walk-in arrivals	beta dist.
service times	log-normal dist.
appointment cancellations	binomial dist.

The intensity (or rate) of the Poission proccess is thus $I_a(c)/364$ per day. Moreover, the duration between two consecutive illness events $e^{\text{ill}}(\rho)$ for patient ρ can be sampled from an exponential distribution with rate $I_a(c)/364$; see Daley and Vere-Jones (2003, Chapter 2).

Type of Acute Illnesses. Whenever an illness event $e^{\text{ill}}(\rho)$ occurs and patient $\rho \in \mathcal{P}$ falls ill, the model generates an acute illness $i \in \mathcal{I}^{\text{act}}$ according to the patients' age class $a \in \mathcal{A}$ and health condition $c \in [0, 1]$. The model assumes a probabilistic link between illness family $f_i \in \mathcal{F}^{\text{act}}$ and the patient's age class a that is expressed via the age class-illness distribution π^{act}; see Section 4.2.5. To that end, any emerging acute illness of patient ρ is randomly assigned to an illness family according to the discrete probability distribution $f \mapsto \pi^{\text{act}}(a, f)$ for $f \in \mathcal{F}^{\text{act}}$.

Qualities of Acute Illnesses. For any new illness $i \in \mathcal{I}^{\text{act}}$ of family $f_i \in \mathcal{F}$ generated through SiM-Care, its seriousness $s_i \in [0, 1]$ depends on a triangular distribution defined on the closed interval $[0, 1]$. The distribution's mode is the health condition $c \in [0, 1]$ of the patient $\rho \in \mathcal{P}$ developing illness i. Thus, patients with a bad health condition tend to develop more serious illnesses.

The duration $d_i \in \mathcal{T}$ of illness i depends on a log-normal distribution. Given i's family of illnesses $f_i \in \mathcal{F}^{\text{act}}$, seriousness $s_i \in [0, 1]$, and the patient's age class $a \in \mathcal{A}$, we define the age-adjusted expected duration of illness i as $\mathbb{E}_a^d(f_i, s_i) \coloneqq \Delta_a^d \cdot D_{f_i}(s_i)$. Therefore, SiM-Care samples the illness' duration d_i from a log-normal distribution with sdlog $\sigma = 0.3$ and meanlog $\mu = \log(\mathbb{E}_a^d(f_i, s_i)) - \sigma^2/2$.

The willingness to wait of patient ρ for the initial treatment of illness i as specified by $\omega_i \in \mathcal{T}$ depends on a Weibull distribution. Given i's family of illness $f_i \in \mathcal{F}^{\text{act}}$, seriousness $s_i \in [0, 1]$, and the developing patient's age class $a \in \mathcal{A}$, the age-adjusted expected willingness to wait of illness i is defined as $\mathbb{E}_a^\omega(f_i, s_i) \coloneqq \Delta_a^\omega \cdot W_{f_i}(s_i)$. Analogous to Wiesche et al. (2017), we sample ω_i from a Weibull distribution with shape parameter $p = 2$ and derive the scale parameter from the age-adjusted expected willingness to wait as $q = \mathbb{E}_a^\omega(f_i, s_i)/\Gamma(1 + (1/p))$ where Γ denotes the gamma function. Figure 4.7 visualizes the resulting density functions for various choices of the age-adjusted expected willingness to wait.

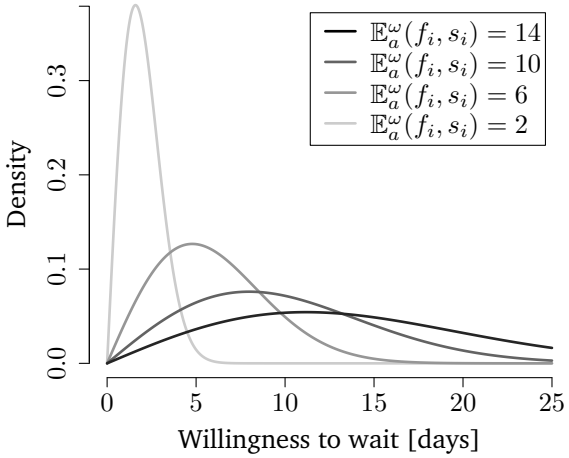


Fig. 4.7.: Weibull distributions of $\omega_i \in \mathcal{T}$ for varying patient's age adjusted expected willingness to wait $\mathbb{E}_a^\omega(f_i, s_i)$.

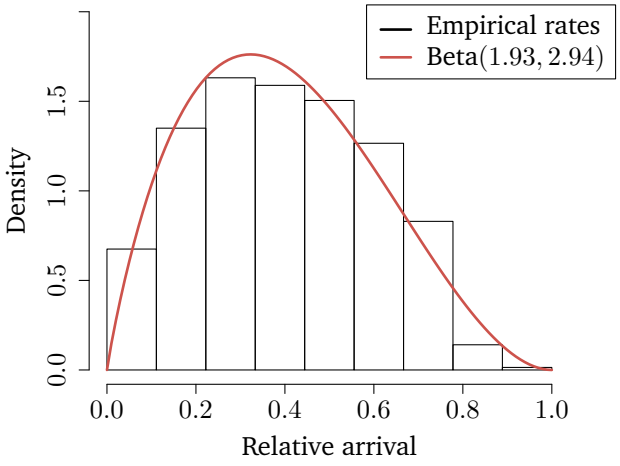


Fig. 4.8.: Histogram and beta distributed maximum-likelihood fit for walk-in arrival rates in Wang et al. (2018).

Patient Punctuality. Patients do not always arrive on time for their scheduled appointments $b \in \mathcal{B}$. Instead, SiM-Care allows for patient arrivals to vary around the scheduled time $t_b \in \mathcal{T}$ of the appointment by including an arrival deviation. As suggested by Cayirli et al. (2006), the arrival deviation from t_b depends on a normal distribution. We choose a mean arrival deviation of $\mu = -5$ minutes and standard deviation of $\sigma = 6$ minutes such that roughly 20% of all patients are expected to arrive late for their appointments which is consistent with the observations reported in Fetter and Thompson (1966).

Walk-in Arrivals. Walk-ins have no prespecified time at which they are expected to arrive. Instead, SiM-Care defines for every walk-in an earliest arrival time $a \in \mathcal{T}$ as well as a latest arrival time $b \in \mathcal{T}$ which are both situational and were thoroughly discussed in Section 4.7.3. The walk-ins' actual arrival within the given feasible arrival interval $[a, b]$ depends on a beta distribution. Specifically, we fit a beta distribution using maximum likelihood estimation to the empirical arrival rates reported by Wang et al. (2018). As a result, we sample the arrival times of walk-ins from the interval $[a, b]$ of feasible arrival times according to a beta distribution with shape parameters $p = 1.93$ and $q = 2.94$; cf. Figure 4.8.

Service Time. SiM-Care treats the service time per patient, i.e., the duration of treatments, as a random parameter. To sample service times, we collected a set of 21 service times in a local primary care practice. As suggested in literature (Wiesche et al., 2017; Cayirli and Veral, 2003), we divide the sample into patients with and without appointment and apply a log-normal maximum likelihood fit. Histograms of our empirical samples and the resulting distributions for patients with appointment and walk-ins are depicted in Figures 4.9 and 4.10, respectively. Based on the fitted distributions, we sample the service times of patients with appointment from a log-normal distribution with meanlog $\mu = 1.82$ and sdlog $\sigma = 0.692$ and the service times for walk-ins from a log-normal distribution with a meanlog $\mu = 1.254$ and

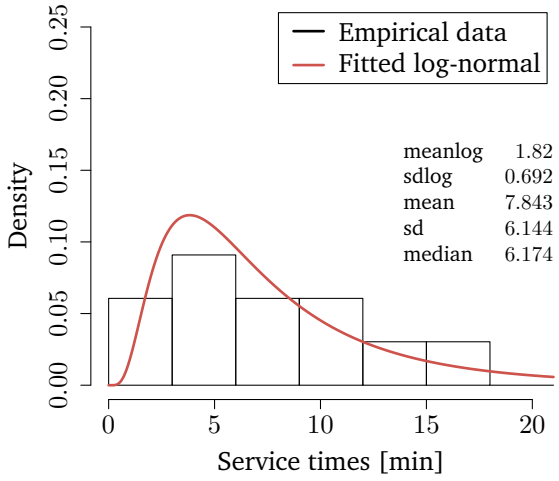


Fig. 4.9.: Histogram and log-normal maximum-likelihood fit for empirical service times of patients with appointment.

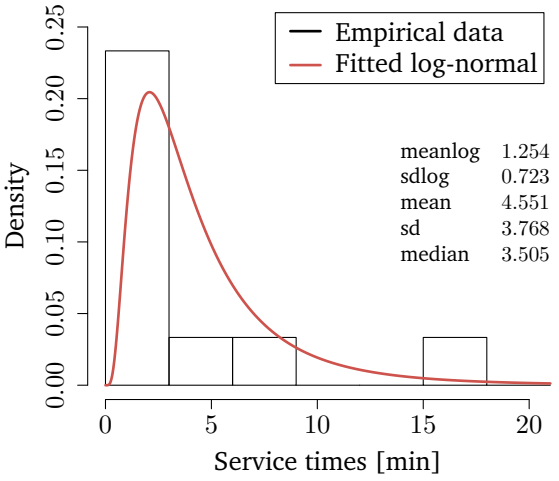


Fig. 4.10.: Histogram and log-normal maximum-likelihood fit for empirical service times of walk-ins.

sdlog $\sigma = 0.723$. As our collected data set does not incorporate transition times, we prolong all sampled service times by one minute.

Appointment Cancellations. Patients that recover from all their current acute illnesses, i.e., $\mathcal{I}^{\text{act}} = \emptyset$, cancel their existing acute appointment $b^{\text{act}} \in \mathcal{B}$ with the age-class specific probability $p_a \in [0, 1]$; compare Section 4.2.4. As long as patients suffer from acute illnesses, they only cancel their acute appointment if they require earlier treatment due to a newly emerged acute illness. All patients that have not canceled their appointment will arrive for it. As chronic illnesses are static within the model, regular appointments are never canceled.

4.5 Emergence and Observation

SiM-Care tracks performance indicators from the point of view of patients, physicians, and policy makers. Thereby, it aims to illustrate the trade-offs between the stakeholders' objectives. As these indicators emerge from agent interactions based on the patients' evolving preferences and the physicians' evolving strategies, they are difficult to predict in general.

From the patients' point of view, performance indicators include access time, access distance, and waiting time. *Access time* measures the time a patient has to wait for an appointment, i.e, given the earliest acceptable appointment time $t \in \mathcal{T}$ and the time of the arranged appointment $t_b \in \mathcal{T}$ it is defined as $\text{ac-time} := t_b - t$. The *access distance* measures the one-way distance patient $\rho \in \mathcal{P}$ has to travel when visiting physician $\phi \in \mathcal{G}$, i.e., $\text{ac-dist} := \text{dist}(\ell_\rho, \ell_\phi)$, where $\text{dist}(\ell_1, \ell_2)$ denotes the driving distance between locations $\ell_1 \in L$ and $\ell_2 \in L$ in kilometers. The patient's *waiting time* measures the time spent on-site before the actual treatment commences. For walk-ins, we define the waiting time for given walk-in arrival $t^{\text{arr}} \in \mathcal{T}$ and treatment commencement $t^{\text{treat}} \in \mathcal{T}$ as $\text{wait-time} := t^{\text{treat}} - t^{\text{arr}}$. For

patients with appointment, we define the waiting time for given time of the appointment $t_b \in \mathcal{T}$, patient's arrival at the practice $t^{\text{arr}} \in \mathcal{T}$, and treatment commencement $t^{\text{treat}} \in \mathcal{T}$ as $\text{wait-time} \coloneqq \max\{t^{\text{treat}} - \max\{t_b, t^{\text{arr}}\}, 0\}$. To evaluate patient's indicators, SiM-Care keeps track of the total access time of arranging acute and regular appointments, the total number of arranged acute and regular appointments, the total number of attended appointments, the total number of walk-ins, the total distance traveled by patients to access physicians, and the total waiting time for both patients with appointment and walk-ins.

From the physicians' point of view, performance indicators include the utilization, overtime, number of treatments, and number of rejected patients with and without appointment. A physician's *utilization* describes the percentage of the available working time spent treating patients, i.e., for a session $\lambda \in \Lambda$ with total treatment duration $t \in \mathcal{T}$ it is defined as $\text{util} \coloneqq t/(\overline{o}(\lambda) - \underline{o}(\lambda) + \frac{1}{24})$. Note that our definition of utilization clearly underestimates a physician's actual utilization as we do not account for additional tasks such as reporting, accounting, and answering phone calls that are not modeled in SiM-Care. *Overtime* describes the physician's working time beyond the anticipated buffer, i.e., if the last patient in session $\lambda \in \Lambda$ is released at time $t^{\text{rel}} \in \mathcal{T}$ it is defined as $\text{over} \coloneqq \max\{t^{\text{rel}} - \overline{o}(\lambda) - \frac{1}{24}, 0\}$; see Figure 4.2. To evaluate the physician's indicators, SiM-Care collects on physician level the total service time spent treating patients, the total number of performed treatments, the total overtime, and the total number of rejected patients with and without appointment. The total available working time per physician that is required to compute the utilization can be derived from the opening hours $o\colon \Lambda/\!\sim \to \mathcal{H} \times \mathcal{H}$ and the modeled time horizon $T \in \mathcal{T}$.

4.6 Input, Initialization, and Warm-Up

SiM-Care codes a large number of values as flexible parameters. Setting up a simulation experiment requires an input scenario to specify these parameter values. Each simulation scenario represents a particular setting, in which a specific set of patients interacts with a specific set of physicians under specific circumstances.

As part of every simulation scenario, the modeler specifies the families of illnesses $\mathcal{F}$, the age classes $\mathcal{A}$, the age class-illness distribution π^{act}, and the set of physicians $\mathcal{G}$ with all their attributes. The set of patients $\mathcal{P}$ is only partially defined through the simulation scenario: Each scenario specifies the total number of chronic and non-chronic patients. Moreover, every patient's location $\ell \in L$, health condition $c \in [0,1]$, age class $a \in \mathcal{A}$, availabilities $\alpha\colon \Lambda/\!\sim \to \{0,1\}$, and (for chronic patients) a chronic illness $\varsigma \in \mathscr{I}^{\text{chro}}$ are given. The remaining attributes of patients, e.g., ratings and acute illnesses, are initialized as described below.

At initialization, patients do not suffer from acute illnesses, i.e., $\mathcal{I}^{\text{act}} = \emptyset$ and are not considered emergencies, i.e., $\varepsilon = 0$. Furthermore, all patients are initialized without scheduled appointments, i.e., $b^{\text{act}} = b^{\text{reg}} = \emptyset$. The consideration set of physicians $\mathcal{G}^{\text{con}} \subseteq \mathcal{G}$ per patient

Algorithm 1: Determine patient's considered PCPs

Input: Patient $\rho \in \mathcal{P}$, set of PCPs $\mathcal{G}$
Output: Patient's consideration set $\mathcal{G}^{\text{con}}$

```
set G^con = ∅
for φ ∈ G do
    if dist(ℓ_ρ, ℓ_φ) < 15 km then
        add φ to G^con
    else
        if rand(20) < 1 then
            add φ to G^con
return G^con
```

$\rho \in \mathcal{P}$ is determined according to Algorithm 1 where rand(x) for $x > 0$ denotes a uniformly distributed float from the half-closed interval $[0, x)$. As a result, each patient considers all physicians within a $15\,\text{km}$ driving radius. Physicians outside this radius are considered with a $5\,\%$ chance as some patients may choose their physician according to criteria other than proximity to their home, e.g., for historical reasons or personal recommendations.

To initialize the appointment ratings $r^{\text{app}}(\phi)$ and walk-in ratings $r^{\text{walk}}(\phi, [\lambda])$ of patient $\rho \in \mathcal{P}$ for every considered physician $\phi \in \mathcal{G}^{\text{con}}$ and weekly session $[\lambda] \in \Lambda/\!\sim$, we denote the number of matches between the physician's opening hours and ρ's availabilities by $m(\rho, \phi) \coloneqq |\{[\lambda] \in \Lambda/\!\sim : \alpha([\lambda]) \wedge o([\lambda]) \neq \emptyset\}|$. Moreover, let $\text{dist}^{\max} \coloneqq \max_{\rho \in \mathcal{P}} \min_{\phi \in \mathcal{G}} \text{dist}(\ell_\rho, \ell_\phi)$ denote the maximal shortest access distance. The model then initializes appointment ratings as

$$r^{\text{app}}(\phi) = \begin{cases} 3\,m(\rho, \phi) - \text{dist}(\ell_\rho, \ell_\phi) + \text{rand}(2\ \text{dist}^{\max}) + 100 & \text{if } m(\rho, \phi) > 0 \\ 0 & \text{else.} \end{cases}$$

Walk-in ratings $r^{\text{walk}}(\phi, [\lambda])$ are session specific as immediate care requires physicians to be in service. Thus, sessions in which a physician is closed are not feasible for walk-in visits which is encoded by an empty rating. The model initializes walk-in ratings as

$$r^{\text{walk}}(\phi, [\lambda]) = \begin{cases} \text{rand}(\text{dist}^{\max}) - \text{dist}(\ell_\rho, \ell_\phi) + 100 & \text{if } o([\lambda]) \neq \emptyset \\ \emptyset & \text{else.} \end{cases}$$

From the initialized ratings, SiM-Care subsequently determines the family physician for chronic patients as the physician from the consideration set that has the highest appointment rating, i.e., $\phi^{\text{fam}} = \text{argmax}_{\phi \in \mathcal{G}^{\text{con}}} r^{\text{app}}(\phi)$ which completes the setup of all simulation entities.

At this point in the initialization process, the global event queue $\mathcal{Q}$ is still empty and therefore running a simulation experiment would result in no agent actions. To make physicians take up their work, the model generates open- and close events $e^{\text{opn}}(\phi)$ and $e^{\text{clo}}(\phi)$ for every

session operated by physician $\phi \in \mathcal{G}$ and adds these to $\mathcal{Q}$. To start the process of patients continuously developing acute illnesses, the model generates an initial illness event $e^{\text{ill}}(\rho)$ for every patient $\rho \in \mathcal{P}$ and adds it to $\mathcal{Q}$. Finally, to start the regular treatments of a patient's chronic illness $\varsigma \in \mathcal{I}^{\text{chro}}$, an initial follow-up event $e^{\text{fol}}(\phi^{\text{fam}}, \rho, \varsigma)$ is generated at a randomly chosen point in time within ς's follow-up interval $\nu_\varsigma \in \mathcal{T}$ according to a uniform distribution and subsequently added to $\mathcal{Q}$.

When creating a simulation experiment in a system such as SiM-Care, modelers can broadly choose one of two approaches: initializing the simulation with an empty system state or with an interim state. An empty system state is inherently unrealistic, as it sets all parameters that are subject to simulation dynamics to zero. For example, an initially empty system state in SiM-Care means that there are no patients with acute illnesses at the start of the simulation. Acute illnesses only arise as simulation time progresses and eventually reach a steady-state when new and subsiding illnesses are in balance. In contrast, initializing a system to an interim state would mean setting all parameters to a value that seems realistic for the simulated system. Revisiting the initialization of acute illnesses, this means that at the start of the simulation, a realistic proportion of patients already suffers from different stages and severities of acute illnesses. Ideally, initializing the system with interim values would mean that there is no period when the system state does not align with the real-world observations. However, such an interim initialization creates additional challenges for validation, as it requires validated values for further parameters. For a structurally valid simulation, these should automatically emerge from an empty-state initialization after a warm-up period. Therefore, we rely on an empty-state initialization and allow for a warm-up period where the simulation state does not align with any plausible real-world state. The duration of the warm-up and the length of the modeled time horizon are both variable and specified by the modeler through the input scenario.

In a further note, this does not correspond to solely analyzing a steady state, as the agents' emergent interactions can result in the development of meaningful trends in the data. For example, even a real-world system may not feature a steady state when it is subject to trends, such as a continuously increasing number of elderly patients, or cycles, such as seasonally changing intensities of certain illnesses. In the examples featured in this thesis, we exclude such trends and cycles and only consider deliberate parameter changes.

4.7 Submodels

We consider different aspects of SiM-Care that rely on an internal logic as submodels. One of the most basic submodels describes the logic of distances and travel times. More complex examples include the logic underlying patients' behavior when requesting appointments and visiting practices as walk-ins, as well as the physician's strategies which are submodels by themselves. As SiM-Care allows for modular PCP strategies, we exemplify each strategy through the specific approach that is used in the case study. Further submodels describe

the consequences of rejecting patients, service time reductions, patients' rating adjustments, patients' choice of their family physician, and treatment effects.

4.7.1 Distances and Travel Times

SiM-Care does not feature a road network to compute travel distances and travel times. Instead, it approximates the driving distance $\mathrm{dist}\colon L \times L \to \mathbb{R}$ between two locations in kilometers using the great circle distance computed through the haversine formula with a detour factor of 1.417 as determined by Boscoe et al. (2012). This publication additionally mentions that driving distances provide good approximations for travel times in minutes, i.e., we compute travel times by assuming a constant driving speed of $60\,\mathrm{km/h}$. As a result we define the travel time $\tau\colon L \times L \to \mathcal{T}$ as $\tau(\ell_1, \ell_2) \coloneqq \frac{\mathrm{dist}(\ell_1,\ell_2)}{60 \cdot 24}$.

4.7.2 Patients Requesting Appointments

Patients $\rho \in \mathcal{P}$ request an appointment with a physician $\phi \in \mathcal{G}$, by specifying the earliest acceptable appointment time $t \in \mathcal{T}$ and their willingness to wait for this appointment $\omega \in \mathcal{T}$. As a result, newly-arranged appointments are feasible, if and only if they are scheduled in the time interval $[t, t + \omega]$.

The earliest acceptable appointment time $t \in \mathcal{T}$ depends on the request. The initial treatment of acute illnesses $i \in \mathcal{I}^{\mathrm{act}}$ is urgent so that patients seek to schedule an appointment as soon as possible. Thus, for these initial treatments, the earliest acceptable appointment time is the time of the request $t^{\mathrm{req}} \in \mathcal{T}$ plus a 30 minute buffer (corresponding to $\frac{1}{48}$ in decimal time) plus the direct travel time, i.e., $t = t^{\mathrm{req}} + \frac{1}{48} + \tau(\ell_\rho, \ell_\phi)$. Follow-up treatments are planned at regular intervals specified by the parameter $\nu_i \in \mathcal{T}$. Patients request follow-up appointments in two ways: First, at the very beginning of the follow-up interval as every patient requests a follow-up appointment directly after the treatment of illnesses that require aftercare. Second, at the very end of the follow-up interval (triggered by a follow-up event) in case no feasible appointment was available at the time of the previous treatment. In the latter case, the request is urgent and therefore the earliest acceptable appointment time is defined as above, i.e., $t = t^{\mathrm{req}} + \frac{1}{48} + \tau(\ell_\rho, \ell_\phi)$. If the follow-up appointment is requested at the beginning of the follow-up interval, the next follow-up appointment for illness $i \in \mathcal{I}$ should be scheduled after the follow-up interval has passed, i.e., we set $t = t^{\mathrm{req}} + \nu_i$.

The willingness to wait $\omega \in \mathcal{T}$ defines the maximum acceptable waiting period between the earliest acceptable appointment time and the actual time of the appointment. As a result, it serves as an upper bound to the patient's access time defined in Section 4.5. Patients' willingness to wait for the initial treatment of acute illness $i \in \mathcal{I}^{\mathrm{act}}$ is illness specific and given by $\omega = \omega_i$. Analogously, the maximum duration chronic patients are willing to wait for their regular appointments depends on their chronic illnesses $\varsigma \in \mathcal{I}^{\mathrm{chro}}$, i.e., $\omega = \omega_\varsigma$. If

patients request a follow-up appointment for acute illness $i \in \mathcal{I}^{\text{act}}$, the willingness to wait is proportional to the length of the follow-up interval $\nu_i \in \mathcal{T}$. To ensure that the follow-up interval is not exceeded by an excessive time span, the willingness to wait for follow-up appointments regarding $i \in \mathcal{I}^{\text{act}}$ is $\omega = \frac{\nu_i}{5} + 1$. Finally, emergency patients who were denied treatment are exceptionally impatient and their willingness to wait is $\omega = 0$.

Algorithm 2 describes how a patient requests an initial appointment for a newly emerged acute illness. First, patients check whether they have a pre-existing appointment within the acceptable time frame. From the patients' point of view, pre-existing appointments are particularly convenient as they require no further actions. Therefore, patients accept pre-existing appointments as feasible, even if they exceed their willingness to wait by up to 12 hours (or $\frac{1}{2}$ in decimal time); see lines $1 - 2$. If the patient's existing appointments are infeasible for the newly emerged illness, the existing acute appointment is canceled to make room for a new, earlier, acute appointment (compare line 4 and 5).

Patients $\rho \in \mathcal{P}$ request appointments from the two currently highest rated physicians $\phi_1, \phi_2 \in \mathcal{G}^{\text{con}}$ in their consideration set (compare line 7). Physicians ϕ_1 and ϕ_2 are queried in order of their rating, i.e., patients first request an appointment with the higher rated PCP ϕ_1 and only resort to ϕ_2 if the request is unsuccessful. In case a physician cannot offer a fitting slot, patients reduce their rating for the respective PCP.

When a patient's willingness to wait is longer than three days (compare line 10), they only accept appointments that fit their personal availability $\alpha \colon \Lambda/\sim \to \{0, 1\}$ (cf. Section 4.2). Otherwise, the request is so urgent that patients are always available.

If neither ϕ_1 nor ϕ_2 offer a feasible slot, the search for a feasible appointment is deemed unsuccessful and patients resort to a walk-in visit.

When patients request follow-up appointments, they mostly follow the steps outlined in Algorithm 2. The main difference concerns the inquiry process (cf. line 7 and 8), as new follow-up appointments are exclusively arranged with the physician that performed the previous treatment. Only pre-existing appointments can be used for follow-up visits although they are not with the physician that performed the previous treatment; compare line 1. If the follow-up appointment request is made at the end of the follow-up interval triggered by a follow-up event, a failure initiates a walk-in attempt to ensure the patient's aftercare. If the follow-up appointment is requested immediately after treatment at the beginning of the follow-up interval, a failure does not lead to a walk-in attempt as the corresponding follow-up event will eventually lead to a reattempt at arranging a follow-up appointment.

Chronic patients' regular appointments are essentially follow-up appointments and thus arranged according to the same logic. The only difference concerns the evaluation of pre-existing appointments. As regular appointments are exclusively arranged with the patient's family physician $\phi^{\text{fam}} \in \mathcal{G}^{\text{con}}$, pre-existing acute appointments are only perceived as feasible if they are with the family physician ϕ^{fam} (cf. line 1 and 2). Infeasible pre-existing acute appointments are not canceled but instead an additional regular appointment is arranged with

Algorithm 2: Arranging appointment for acute illness

Input: Patient $\rho \in \mathcal{P}$, willingness to wait $\omega \in \mathcal{T}$, earliest appointment time $t \in \mathcal{T}$
Output: Was a feasible appointment found or arranged?

1 **if** *ρ has acute or regular appointment before time $t + \omega + \frac{1}{2}$* **then**
2 **return** True
3 **else**
4 cancel acute appointment
5 delete associated arrival event $e^{\text{arv}}(\phi, \rho)$
6 determine preferred physicians $\phi_1, \phi_2 \in \mathcal{G}^{\text{con}}$ such that $r^{\text{app}}(\phi_1) \geq r^{\text{app}}(\phi_2) \geq r^{\text{app}}(\phi)\ \forall \phi \in \mathcal{G}^{\text{con}} \setminus \{\phi_1, \phi_2\}$
7 **for** $j = 1, 2$ **do**
8 query ϕ_j for an appointment
9 **if** *physician ϕ_j offers appointment within $[t, t + \omega] \wedge$ (satisfying ρ's availabilities $\alpha \vee \omega \leq 3$)* **then**
10 accept appointment # adapt $r^{\text{app}}(\phi_j)$
11 add $e^{\text{arv}}(\phi_j, \rho)$ to $\mathcal{Q}$
12 **return** True
13 **else**
14 refuse appointment # adapt $r^{\text{app}}(\phi_j)$
15 **continue**
16 **return** False

the family physician ϕ^{fam} (cf. line 4 and 5). Only if the newly arranged regular appointment is before or at most 12 hours after an existing acute appointment, i.e., $t_{b^{\text{reg}}} \leq t_{b^{\text{act}}} + \frac{1}{2}$, the latter is canceled as all acute illnesses will be treated at the regular appointment.

4.7.3 Walk-in Decision Making

Within SiM-Care, all walk-in visits are preceded by an unsuccessful appointment request. As walk-in visits are per se urgent, the earliest possible time $t \in \mathcal{T}$ for a walk-in visit of patient $\rho \in \mathcal{P}$ at physician $\phi \in \mathcal{G}^{\text{con}}$ is, analogous to Section 4.7.2, defined as the current time $t^{\text{curr}} \in \mathcal{T}$ plus a 30 minute buffer plus the direct travel time, i.e., $t = t^{\text{curr}} + \frac{1}{48} + \tau(\ell_\rho, \ell_\phi)$. The patients' willingness to wait for the walk-in visit is defined as the willingness to wait $\omega \in \mathcal{T}$ of the preceding appointment request. As a result, the patient's walk-in visit takes place in the time interval $[t, t + \omega]$, unless this is impossible due to the physicians' opening hours.

As part of the walk-in decision making, patients decide on a physician $\phi^* \in \mathcal{G}^{\text{con}}$ and session $\lambda^* \in \Lambda$ for their walk-in visit. To that end, SiM-Care computes all physician-session combinations $W \subseteq \mathcal{G}^{\text{con}} \times \Lambda$ that fall into the interval $[t, t + \omega]$ and thus can be targeted for a walk-in visit. If $W = \emptyset$, the model gradually increases the willingness to wait ω until $W \neq \emptyset$.

Patients select the physician-session combination $(\phi^*, \lambda^*) \in W$ targeted for their walk-in visit on the basis of their walk-in ratings r^{walk} via

$$(\phi^*, \lambda^*) = \text{argmax}_{(\phi,\lambda)\in W}\, 0.95^{w_t(\lambda)} \cdot r^{\text{walk}}(\phi, [\lambda]),$$

where $w_t(\lambda) \coloneqq \overline{o}(\lambda) - t$ denotes the time difference between the earliest possible walk-in time $t \in \mathcal{T}$ and the end of session $\lambda \in \Lambda$. This takes into account that walk-ins urgently want to visit a physician by discounting the ratings based on the approximate access time $w_t(\lambda) \geq 0$. Note that this discounting model yields undesired results if we allow for negative ratings, motivating the models limitation to non-negative ratings.

Given the targeted physician-session combination $(\phi^*, \lambda^*) \in W$ for the walk-in visit, the time interval during which the actual visit at ϕ^* may take place is defined as follows. The earliest time for walk-ins to arrive in session $\lambda^* \in \Lambda$ is 15 minutes (or $\frac{1}{96}$ in decimal time) before its beginning $\underline{o}(\lambda^*) \in \mathcal{T}$, but obviously not before the earliest possible arrival $t \in \mathcal{T}$. The latest possible arrival in session $\lambda^* \in \Lambda$ is its ending $\overline{o}(\lambda^*) \in \mathcal{T}$, but not after the latest possible arrival $t + \omega$. The resulting time interval for the patient's walk-in arrival is

$$\left[\max(\underline{o}(\lambda^*) - \tfrac{1}{96},\ t),\ \min(\overline{o}(\lambda^*), t + \omega)\right].$$

The patient's actual arrival within the feasible time interval is stochastic and sampled according to the distribution specified in Section 4.4.

As long as patients actively pursue walk-in treatment, they never arrange new appointments. That is, if a walk-in develops a new acute illness or seeks an immediate follow-up appointment triggered by a follow-up event, their need for medical attention is met through the ongoing walk-in visit.

4.7.4 Service Time Reduction

Physicians' treatment strategies let them reduce service times to prevent congestion and minimize overtime. Within the model, the service time reduction operationalizes via a multiplicative factor $\zeta \in [0, 1]$. Thus, a treatment with an original service time of 10 minutes (sampled from the log-normal distribution described in Section 4.4) takes only 8 minutes when performed by a physician with current consultation speed $\zeta = 0.8$. When there is no effort to reduce service times, i.e., $\zeta = 1$, the actual services time coincide with the sampled original service times.

4.7.5 Consequences from Rejection of Patients

Whenever a patient visits a physician either with an appointment or as a walk-in, the physician's admission strategy determines whether the patient is admitted or rejected. Following

Tab. 4.7.: Adaptation of patient ratings r^{app} and r^{walk}, where $\omega \in \mathcal{T}$ describes patient's willingness to wait and $\zeta \in [0, 1]$ the physician's consultation speed.

Positive Event	Adjust
waiting time $< 7\,\text{min}$	$+5$
successful arrangement of appt.	$+4$
successful treatment as walk-in	$+3\zeta$
successful treatment with appt.	$+2\zeta$

Negative Event	Adjust
waiting time $> 30\,\text{min}$	-10
no appt. within willingness available	$-\omega$
rejected as walk-in	-10
rejected with appt.	-20

a rejection, patients reduce their personal ratings r^{app} or r^{walk} depending on whether they arrived for an appointment or as a walk-in. As rejected patients have been denied treatment, they are subsequently flagged as emergencies, i.e, $\varepsilon = 1$. In order to be treated, rejected patients then start a walk-in attempt with reduced willingness to wait $\omega = 0$, i.e., they visit their preferred physician according to the updated walk-in preferences r^{walk} in the earliest possible session; compare Section 4.7.3. A patient's emergency flag is only revoked after the next successful treatment or if the patient fully recovers from all acute illnesses.

4.7.6 Rating Adjustments

Throughout the simulation, patients adjust their ratings of physicians according to their experiences via additive factors. To that end, patients increase ratings based on positive experiences and decrease ratings following negative experiences. Thereby, patients with an appointment update their appointment ratings r^{app} while walk-ins update their walk-in ratings r^{walk}. Table 4.7 lists all events that trigger a rating adjustment.

In SiM-Care, only the effect of a failed appointment request and the effect of a successful treatment are parameterized. All other event effects are hard-coded to represent the following intuition about patient perceptions: Unanticipated events cause a stronger adjustment, while anticipated events only cause a slight adjustment. For example, visiting a physician with an appointment and not being admitted is considered highly unlikely and therefore highly penalized. Furthermore, patients react more strongly to negative experiences, reflecting the so-called *negativity bias* (Baumeister et al., 2001).

In case a physician fails to offer a fitting appointment, the negative adjustment depends on the patient's associated willingness to wait $\omega \in \mathcal{T}$. As $\omega \geq 0$, the adjustment $-\omega$ is always non-positive. If the willingness to wait is high, the expectation of receiving a fitting slot is also high, so that the resulting disappointment leads to a stronger negative adjustment.

When physicians reduce their service time as part of their treatment strategy, patients feel rushed. Therefore, the model scales the positive adjustment following a successful treatment as dependent on the physician's current consultation speed. For example, at a consultation speed of $\zeta = 0.5$ a successful treatment with appointment increases r^{app} only by a value of $0.5 \cdot 2 = 1$; compare Table 4.7.

Algorithm 3: Reevaluation of family physician

Input: Chronic patient $\rho \in \mathcal{P}$
Output: Family physician ϕ^{fam}

let $\phi^* = \text{argmax}_{\phi \in \mathcal{G}^{\text{con}}} r^{\text{app}}(\phi)$
if $r^{app}(\phi^*) \geq 1.2 \cdot r^{app}(\phi^{fam})$ **then**
    $\phi^{\text{fam}} = \phi^*$
return ϕ^{fam}

To ensure the desired behavior of discounting ratings as described in Section 4.7.3, we bound all ratings from below by zero, i.e., we enforce $r^{\text{app}}(\phi) \geq 0$ and $r^{\text{walk}}(\phi, [\lambda]) \geq 0$ for all $\rho \in \mathcal{P}$, $\phi \in \mathcal{G}$, and $[\lambda] \in \Lambda/\sim$. As a result, negative adjustments have no effect on physicians with rating zero.

4.7.7 Family Physician Adjustments

Every time chronic patients adjust their appointment ratings $r^{\text{app}}(\phi)$ for $\phi \in \mathcal{G}^{\text{con}}$, they simultaneously reevaluate their family physician ϕ^{fam} according to Algorithm 3. Thereby, chronic patients change their family physician as soon as another physician from the consideration set has a rating that is at least $20\,\%$ higher than the current family physician's rating.

4.7.8 Treatment Effects

Physicians treat all of a patient's current acute illnesses $i \in \mathcal{I}^{\text{act}}$ during the same appointment. As a result, all scheduled follow-up events $e^{\text{fol}}(\phi, \rho, i)$ for $i \in \mathcal{I}^{\text{act}}$ are deleted. Moreover, all illnesses $i \in \mathcal{I}^{\text{act}}$ that require only a single treatment, as indicated by $d_i = \emptyset$, are cured and thus removed from $\mathcal{I}^{\text{act}}$. Finally, new follow-up events are scheduled for all illnesses $i \in \mathcal{I}^{\text{act}}$ that still require follow-up consultation as indicated by a positive follow-up interval $\nu_i > 0$.

Chronic illnesses are only treated during the recurrent regular appointments or during walk-in visits triggered by the unavailability of a feasible regular appointment. If $\varsigma \in \mathcal{I}^{\text{chro}}$ is treated, any existing follow-up event $e^{\text{fol}}(\phi, \rho, \varsigma) \in \mathcal{Q}$ is deleted and replaced by a new, updated one.

Finally, the successful treatment revokes any emergency flag the patient may have, i.e., we set $\varepsilon = 0$.

4.7.9 PCP Strategies

PCP strategies determine physicians' decision making through exchangeable submodels that are defined as part of every scenario. For illustration, we describe the exemplary strategies implemented and evaluated in our case study.

Appointment scheduling strategy. *Individual-block/ Fixed-interval* (IBFI) evenly spaces out appointments throughout each session; see Cayirli and Veral (2003) and Klassen and Rohleder (1996). To that end, it divides the opening hours of each session in a 140 day rolling horizon into slots of 15 minutes length. Each slot can accommodate one appointment and slots are offered to patients on a first-come-first-served (FCFS) basis. Thus, no appointments are withheld and every patient is offered the earliest feasible appointment at the time of inquiry.

Treatment strategy. *Priority first come, first served* (PFCFS) is popularly used in studies of health systems (Cayirli and Veral, 2003). In PFCFS, patients with appointment are prioritized over walk-ins and within their respective groups, patients are served in order of their arrival, i.e., FCFS; compare Rising et al. (1973) and Cox et al. (1985). Patients that arrive before the beginning $\underline{o}(\lambda) \in \mathcal{T}$ of session $\lambda \in \Lambda$ have to wait and the physician does not start treatments until the session has officially begun. The PCP's standard consultation speed in PFCFS is $\zeta = 1.0$, which is adjusted to $\zeta = 0.8$ whenever more than three patients await treatment; compare Section 4.7.4.

Admission strategy. *Priority threshold* (PT) admits patients up to a certain utilization threshold; compare Kim et al. (2015) and Qu et al. (2015). PT differentiates between appointment, walk-in, and emergency patients. Emergency patients are always admitted, i.e., they have an infinite admission threshold. Patients with an appointment in session $\lambda \in \Lambda$ are admitted as long as their time of arrival $t^{\text{arr}} \in \mathcal{T}$ is before the end of the session's buffer, i.e., $t^{\text{arr}} \leq \overline{o}(\lambda) + \frac{1}{24}$. Appointment patients that arrive after the session's anticipated buffer are rejected. For the admittance of walk-ins, physicians predict their remaining workload by multiplying an expected service time with the number of currently waiting patients and upcoming scheduled appointments. If this estimated workload is lower than the remaining duration of the current session including buffer, walk-ins are admitted, otherwise rejected. The expected service time is initialized to 7 minutes and adjusted at the end of each session as follows. On the one hand, the expected service time is increased by one minute if three or more patients are awaiting treatment at the end of the anticipated buffer. On the other hand, the expected service time is reduced by 20 seconds if the physician is idle at the end of the buffer although walk-ins were previously rejected.

4.8 Structural Validation and Verification

In SiM-Care, validation and verification were carried out according to the best practices documented in the literature (Kleijnen, 1995; Sargent, 2013). To ensure that our model implementation is correct (verification), we followed established good programming practices. That is, we used object oriented programming to write modular code. SiM-Care is implemented in Java using OpenJDK 11 (Oracle, 2018). All random distributions are

implemented using the Apache Common Math library (Math Commons, 2016). Each module is individually verified through unit testing. Assertions ensure that variables remain within their specifications at runtime. As an additional mean to detect undesired model behavior, SiM-Care can trace the entire simulation process. Traces are specialized logs that contain all information about the model's execution. In SiM-Care, traces are textual and comprehensible to modelers. They enable the tracking of agents through the overall model and contain all the information that would be required to animate the model. Analyzing traces and input output relationships, we performed dynamic tests for multiple simulation scenarios of various sizes with different system setups.

To ensure that our conceptional model serves as an adequate representation of real primary care systems (validation), we took several measures. With regard to face validity, we presented the conceptual model to physicians and decision makers from health insurers as well as public authorities. Furthermore, SiM-Care builds on data from the literature as well as empirical data collected on-site. Moreover, we visited a primary care practice and interviewed staff to capture and understand the daily processes and routines of PCPs. For the specific scenarios featured in the case study, we validated the simulation output with available empirical data. Details on this historical validation can be found in the baseline analysis of the following case study.

5 Case Study: Effects of Demographic Change

To demonstrate the potential of SiM-Care, we present a case study evaluating the effects of changes in the population of a primary care system. Specifically, we create a baseline scenario representing a real-world primary care system in the district of Aachen and investigate two possible changes in the primary care system's population from the status quo. On the one hand, a decline in the number of PCPs as a result of a decreasing interest in opening a primary care practice in rural areas; see Jacob et al. (2015). On the other hand, an aging of the population causing a shift in the quality and intensity of illnesses and the resulting health care requirements. By considering both of these changes individually and in combination, we create three "what-if" scenarios that we compare to the baseline scenario.

Each scenario models a time period of one year preceded by warm-up period. As SiM-Care relies on stochastic values, every simulation experiment includes 20 independent runs. Section 5.1 details how the baseline scenario is derived from empirical data. Section 5.2 documents the analysis and validation of the baseline scenario. Sections 5.3, 5.4, and 5.5 describe how the considered changes in the three "what-if" scenarios are implemented in SiM-Care and subsequently benchmark these against the baseline scenario. Finally, a partial sensitivity analysis is provided in Section 5.6.

5.1 Baseline Scenario

The real-world primary care system that serves as the template for our study comprises three predominantly rural municipalities (Roetgen, Simmerath, and Monschau) in Western Germany with a total population of approximately $35{,}000$ inhabitants and 20 PCPs. In order to capture a real-world primary care system in the form of a simulation scenario, empirical data is required. Most of this data is specific to a primary care system or its country of origin such that data collection has to be carried out for each system individually. For the considered primary care system, empirical data concerning the physicians' distribution and opening hours was provided by the responsible department of public health or obtained from the responsible association of statutory health insurance physicians (Kassenärtzliche Vereinigung Nordrhein, 2019). The distribution of patients and their demographic composition is available from the national census (Information und Technik Nordrhein-Westfalen, 2016) and official population projections by the federal state (Information und Technik Nordrhein-Westfalen, 2019). The distribution of illnesses and their characteristics can be estimated from publications of health

Tab. 5.1.: Basis for the selection of input parameters.

Attribute	Basis (Source)
PCPs	
location	empiric (department of public health)
opening hours	empiric (Kassenärtzliche Vereinigung Nordrhein, 2019)
strategies	literature (Cayirli and Veral, 2003; Klassen and Rohleder, 1996; Kim et al., 2015)
Patients	
location	empiric (Information und Technik Nordrhein-Westfalen, 2016)
age class	empiric (Information und Technik Nordrhein-Westfalen, 2016)
health condition	inferred
Age classes	
exp. annual acute illnesses	inferred
dev. illness duration	inferred
dev. willingness to wait	inferred
availabilities	inferred
appointment cancellation	inferred
chronic patients	empiric (Robert Koch-Institut, 2014)
Families of Illnesses	
characteristics	inferred
age class-illness distribution	empiric (Grobe et al., 2011)

insurances and federal government agencies (Grobe et al., 2011; Robert Koch-Institut, 2014). All unavailable data was either empirically collected in a primary care practice or, where this was not possible, inferred. For the sake of clarity, we summarize our basis for the selection of each input parameter in Table 5.1.

In the following, we discuss how the available empirical data translates into a simulation scenario. To that end, we detail the input parameter choices, i.e., the modeled physicians, patients, age classes, families of illnesses, and age class-illness distributions.

5.1.1 Primary Care Physicians

The population of physicians $\mathcal{G}$ in our baseline scenario aims to model the actual physicians in the considered primary care system. According to data provided by the Aachen department of public health in 2017, there are 20 PCPs with health insurance accreditation in the three municipalities under consideration. The physicians' exact locations are specified as part of the provided dataset (cf. Figure 5.1) and the physicians opening hours were obtained from the Association of Statutory Health Insurance Physicians Nordrhein (Kassenärtzliche Vereinigung Nordrhein, 2019). Remark that according to these opening hours, all considered physicians are closed on Saturdays and Sundays. Concerning the employed strategies, all physicians $\phi \in \mathcal{G}$ apply the IBFI appointment scheduling strategy, PFCFS treatment strategy, and PT admission strategy; cf. Section 4.7.9.

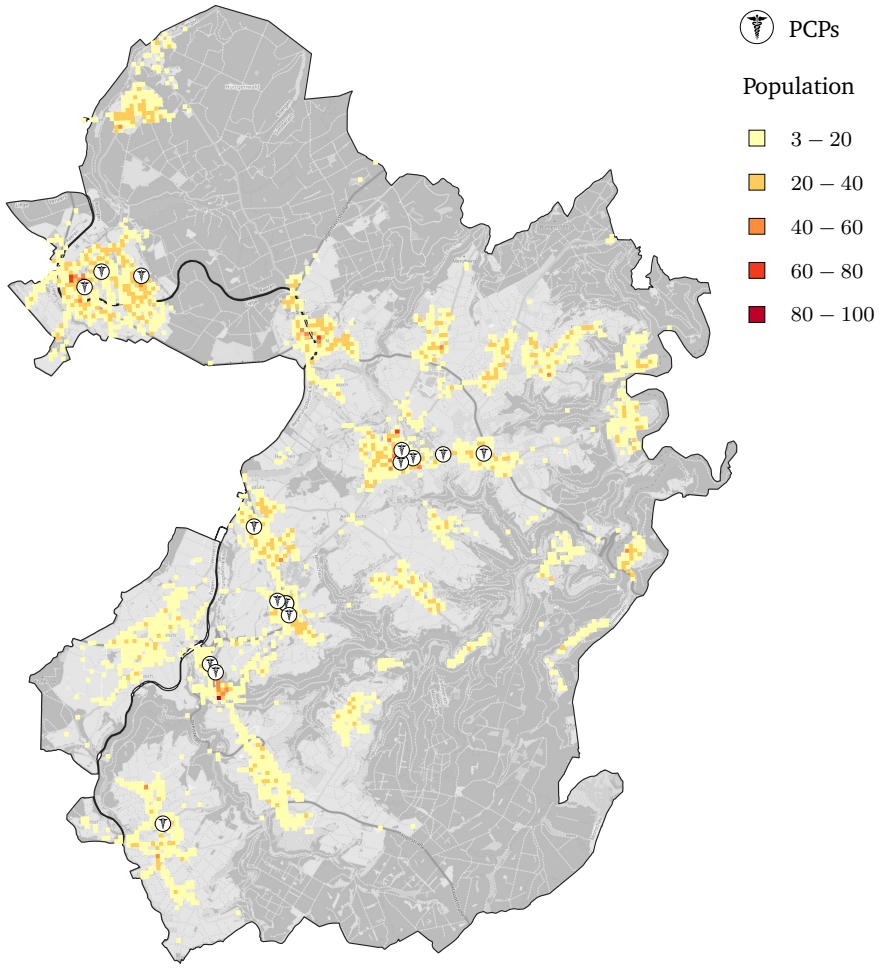


Fig. 5.1.: Locations of PCPs with health insurance accreditation and population cells reported by the 2011 census (Information und Technik Nordrhein-Westfalen, 2016).[2]

5.1.2 Patients

The population of patients $\mathcal{P}$ in the baseline scenario aims to reflect the actual population in the considered primary care system. The latest publicly available high resolution population data for the considered region is the German census conducted in 2011 (Information und Technik Nordrhein-Westfalen, 2016). At a resolution of 2,754 population cells measuring one hectare each, the 2011 census reports a total population of 35,542 for the three municipalities Roetgen, Simmerath, and Monschau; compare Figure 5.1. This population includes children under the age of 16 who are excluded from our considerations, as children mainly consult pediatricians who are not modeled in this study. Census data does not state the exact number of under 16-year-olds in each population cell. Instead, the census reports the total number of under 16-year-olds on municipality level: Roetgen 1,390, Simmerath 2,383, and Monschau 1,794. To exclude children under the age of 16, we proceed as follows. First, we fix one adult per population cell as we assume that children under the age of 16 do not live on their own. Then, we sample the number of under 16-year-olds from the remaining population of each

[2]Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

municipality according to a uniform distribution. Exemplifying this procedure for Roetgen, the census reports a total population of $8{,}288$ distributed over 534 population cells. We fix one adult per population cell, and uniformly distribute the $1{,}390$ under 16-year-olds among the $7{,}754$ remaining inhabitants. Performing this procedure for each municipality individually, we obtain the final patient population $\mathcal{P}$ consisting of $29{,}975$ patient agents distributed over $2{,}754$ population cells.

The age-class independent attributes of each patient $\rho \in \mathcal{P}$ are determined as follows. The location $\ell \in L$ for each patient is sampled from the associated population cell according to a uniform distribution. Patients' health conditions $c \in [0, 1]$ are sampled from a beta distribution with shape parameters $p = q = 25$ such that all patients have an expected health condition of $\mathbb{E}(c) = 0.5$.

5.1.3 Age classes

SiM-Care accounts for the age dependency of various patient characteristics through the concept of age classes. The baseline scenario differentiates three patient age classes: young (16-24), middle-aged (25-65), and elderly (>65). The characteristics of the modeled age classes $\mathcal{A}$ are shown in Table 5.2. Young patients (16-24) are, on average, the healthiest among all patients. Thus, they are expected to develop the fewest acute illnesses per year from which they recover relatively quickly. Their expected willingness to wait is prolonged and they are very unlikely to visit a PCP unless it is necessary. Middle-aged patients (25-65) represent the working share of the population and we consider them to be our "nominal" patients. They thus do not deviate from the expected illness duration and the expected willingness to wait as specified by families of illnesses. On average, middle-aged patients (25-65) develop more acute illnesses per year than young patients (16-24) while keeping slightly more appointments after recovery. Elderly patients (>65) are expected to develop the most annual acute illnesses and it takes them more time to recover from these. Their expected willingness to wait is the lowest among all age classes and they are most likely to visit a PCP after all symptoms have subsided.

Based on census data (Information und Technik Nordrhein-Westfalen, 2016), the age class $a \in \mathcal{A}$ of each patient depends on the discrete probability distribution shown in Table 5.3. The age-class dependent attributes of each patient $\rho \in \mathcal{P}$ are subsequently determined as follows. Each patient's session availabilities α are determined by performing a Bernoulli trial for every session of the week $[\lambda] \in \Lambda/\sim$ based on the age-class dependent success probabilities from Table 5.3. To decide whether a patient is chronically ill, we perform a Bernoulli trial using the age-class dependent success probabilities from Table 5.3 that were estimated based on Robert Koch-Institut (2014).

Tab. 5.2.: Characteristics of considered age classes $a \in \mathcal{A}$.

	16-24	25-65	>65
exp. illnesses	$I_a(c)$=$6c$	$I_a(c)$=$7c$+1	$I_a(c)$=$9c$+1
dev. duration	Δ_a^d=0.8	Δ_a^d=1.0	Δ_a^d=1.2
dev. willingness	Δ_a^ω=1.2	Δ_a^ω=1.0	Δ_a^ω=0.8
prob. appt. cxl.	p_a=0.95	p_a=0.8	p_a=0.7

Tab. 5.3.: Age specific parameters for patient generation.

	16-24	25-65	>65
age class distribution	0.1196	0.6318	0.2486
availability probability	0.85	0.55	0.95
chronic illness probability	0.12	0.33	0.52

5.1.4 Families of Illnesses

The most important classification system for illnesses world-wide is the International Statistical Classification of Diseases and Related Health Problems (ICD) maintained by the World Health Organization. In its current revision, ICD-10 (World Health Organization, 2004) distinguishes more than $14{,}000$ codes. For the purpose of SiM-Care, such a granular illness distinction is generally not necessary. Thus, we can aggregate ICD-10 codes, e.g., using the 22 chapters of ICD-10, or considering only a subset of all ICD-10 codes, e.g., the ones most frequently reported. In the baseline scenario, we consider a subset of the 100 ICD-10 codes most frequently reported to the Association of Statutory Health Insurance Physicians Nordrhein (Kassenärtzliche Vereinigung Nordrhein, 2018). The attributes of families of illnesses can be estimated based on historical treatment data which is commonly available to health insurers. Yet, such data is naturally protected by confidentiality and cannot be published. Thus, we choose a less elaborate approach and only estimate all attributes which yields the families of illnesses $\mathcal{F}$ listed in Table 5.4.

5.1.5 Age Class-Illness Distributions

Age class-illness distributions define the expected occurrence of acute families of illnesses $f_i \in \mathcal{F}^{\text{act}}$ per age class $a \in \mathcal{A}$. For this distribution, the baseline scenario relies on the reported incidence rates of 8.2 million customers of a large German health insurer published in Grobe et al. (2011). We aggregate this data by gender and age to obtain the age class-illness distribution $\pi^{\text{act}} \colon \mathcal{A} \times \mathcal{F}^{\text{act}} \to [0,1]$ shown in Table 5.5. Analogously, we determine the expected distribution of chronic families of illnesses $\mathcal{F}^{\text{chro}}$ among the modeled age classes $\mathcal{A}$ denoted by $\pi^{\text{chro}} \colon \mathcal{A} \times \mathcal{F}^{\text{chro}} \to [0,1]$ shown in Table 5.5.

The distribution π^{chro} is not part of the baseline scenario itself. Instead, it is only required to generate the unique chronic illness of chronically ill patients. In the baseline scenario, we generate every chronic patient's chronic illness $\varsigma \in \mathscr{I}^{\text{chro}}$ analogously to the process of generating acute illnesses as described in Section 4.4. Given the patient's age class $a \in \mathcal{A}$, the illness family $f_\varsigma \in \mathcal{F}^{\text{chro}}$ of ς depends on the discrete probability distribution $f \mapsto \pi^{\text{chro}}(a, f)$ for $f \in \mathcal{F}^{\text{chro}}$. The seriousness $s_\varsigma \in [0,1]$ of ς is sampled from a triangular distribution using

Tab. 5.4.: Characteristics of considered families of illnesses $f \in \mathcal{F}$.

ICD	Name	Exp. willingness	Exp. duration	Treatment freq.	chronic
I10	high blood pressure	$W_f(s) = -10s+20$	not applicable	$N_f(s) = -20s+100$	True
E11	diabetes	$W_f(s) = -4s+14$	not applicable	$N_f(s) = -10s+90$	True
I25	ischemic heart disease	$W_f(s) = -4s+10$	not applicable	$N_f(s) = -30s+100$	True
E78	high cholesterol level	$W_f(s) = -5s+8$	$D_f(s) = 4s+8$	$N_f(s) = -2s+11$	False
M54	back pain	$W_f(s) = -3s+4$	$D_f(s) = 9s+5$	$N_f(s) = -4s+11$	False
Z25	vaccination	$W_f(s) = 40$	not applicable	not applicable	False
J06	cold	$W_f(s) = -2s+2$	$D_f(s) = 5s+4$	$N_f(s) = -s+6$	False

Tab. 5.5.: Age class-illness distributions π^{act} for acute illnesses and π^{chro} for chronic illnesses.

	16-24	25-65	>65
high cholesterol level	0.02	0.24	0.36
back pain	0.32	0.38	0.28
vaccination	0.14	0.14	0.27
cold	0.52	0.24	0.09
high blood pressure	0.17	0.65	0.61
diabetes	0.33	0.16	0.2
ischemic heart disease	0.5	0.19	0.19

the patient's health condition $c \in [0, 1]$ as mode. In turn, the seriousness defines treatment frequency of ς via $\nu_\varsigma = N_{f_\varsigma}(s_\varsigma)$ and willingness to wait as $\omega_\varsigma = W_{f_\varsigma}(s_\varsigma)$.

5.1.6 Duration of Warm-up

Every run of SiM-Care contains a warm-up period; compare Section 4.6. To determine an appropriate length for the warm-up period, we simulate the baseline scenario for a time period of 70 years and track all performance indicator for each year individually. Figure 5.2 shows the resulting evolution for the average access time of appointments, average daily overtime of physicians, and average waiting time of walk-ins. As we can see, all performance indicators are evolving in the first 30 to 50 years before they stabilize. Similar behaviors can be observed for all other measured performance indicators. Emphasizing long-term stability, we set the duration of the warm-up period in each scenario to 60 years and note that a warm-up duration half as long seems also valid.

5.2 Baseline Analysis

Table 5.6 reports the resulting mean performance indicators as well as the associated exact $95\,\%$-confidence intervals for each tracked performance indicator; compare Section 4.5. The results show that in the status quo, each physician in our primary care system performs, on average, $10{,}122.16$ treatments per year. This amounts to an average number of 6.75 physician contacts per patient which is slightly above the 6.6 annual PCP contacts reported back in 2006 (Aho, 2006). Roughly $47\,\%$ of patients visiting a physician in our baseline scenario are walk-ins, which is consistent with the observed $48\,\%$ share of walk-ins in our

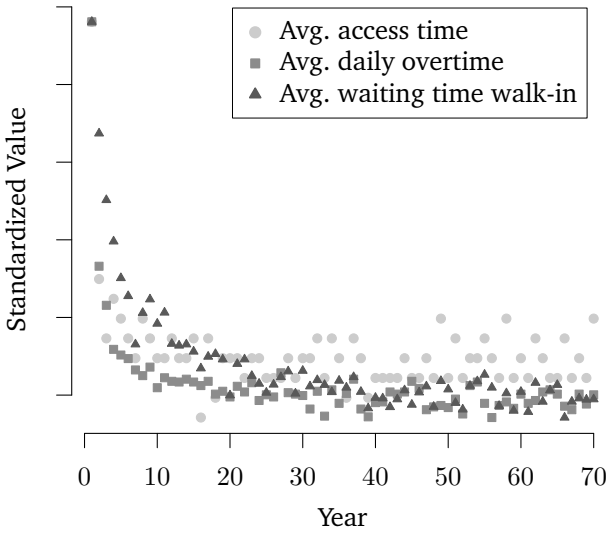


Fig. 5.2.: Evolution of performance indicators in the baseline scenario for every year in a period of 70 years.

Tab. 5.6.: Mean performance indicators and 95 %-confidence intervals obtained by repeating each simulation experiment 20 times for the baseline scenario.

	Baseline Scenario	
	Mean	95 %-CI
average number of treatments	10,122.16	[10,112.1, 10,132.22]
average number of walk-ins	4,731.53	[4,721.61, 4,741.46]
average number of acute appointments	3,215.59	[3,214.14, 3,217.05]
average number of regular appointments	2,175.03	[2,173.6, 2,176.47]
average utilization [%]	72.15	[72.08, 72.22]
average daily overtime [min]	0.8	[0.74, 0.86]
average number of rejected walk-ins	13.85	[13.16, 15.53]
average access time [d]	2.46	[2.45, 2.47]
average access time regular [d]	1.49	[1.48, 1.51]
average access distance [km]	4.95	[4.94, 4.96]
average waiting time appt. [min]	2.09	[2.08, 2.1]
average waiting time walk-in [min]	39.75	[39.64, 39.85]
on-time appointments [%]	61.13	[61.02, 61.24]
number of acute illnesses	136,454.2	[136,283.89, 136,624.52]
number of chronic patients	10,662	–
total PCP capacity [h]	32,617	–

collected empirical dataset of service times; compare Section 4.4. Concerning overtimes, we were unable to obtain empirical data as most PCPs are self-employed and even the definition of overtime is unclear. However, the estimated average daily overtime per physician (according to our definition) seems to be too low at just 0.8 minutes per day. This can be explained by the fact that we incorporate buffers at the end of each session and do not include additional mandatory physician's activities such as reporting and accounting into our simulation model.

Patients in our baseline scenario are expected to travel almost 5 km to visit a physician and have to wait an average number of 2.46 days for their appointments. With regard to waiting times, we obtain an average expected waiting time of 2.09 minutes for patients with appointment and 39.75 minutes for walk-ins. In comparison to the average waiting times

Tab. 5.7.: Populations in each simulation scenario variant.

	Decl. PCPs Scen. 1		Aging Patients Scen. 2		Comb. Effects Scen. 3	
	s	m	s	m	s	m
patients	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}^{\mathrm{s}}$	$\mathcal{P}^{\mathrm{m}}$	$\mathcal{P}^{\mathrm{s}}$	$\mathcal{P}^{\mathrm{m}}$
PCPs	$\mathcal{G}^{\mathrm{s}}$	$\mathcal{G}^{\mathrm{m}}$	$\mathcal{G}$	$\mathcal{G}$	$\mathcal{G}^{\mathrm{s}}$	$\mathcal{G}^{\mathrm{m}}$

s = short-term shift, m = medium-term shift

observed when recording our service time dataset (4 minutes with appointment, 15 as walk-in), our simulated waiting times are strikingly unfavorable for walk-ins which suggests that physicians avoid excessive waiting times of walk-ins through more sophisticated treatment strategies, e.g., accumulating priority queues (Stanford et al., 2014).

5.3 Scenario 1: Decline in PCPs

Scenario 1 models a decline in the number of PCPs for a short- and a medium-term shift in time. To that end, we exclude all physicians from our baseline PCP population $\mathcal{G}$ that reached the statutory retirement age of 65 by this point. Specifically, we consider the year 2023 by which 4 out of 20 PCPs will have reached the statutory retirement age as well as the year 2027 by which 7 out of 20 PCPs will have reached the statutory retirement age. Assuming that none of the excluded physicians are replaced by a successor, we obtain our decimated population of physicians $\mathcal{G}^{\mathrm{s}}$ for the short-term and $\mathcal{G}^{\mathrm{m}}$ for the medium-term shift. By replacing the physician population $\mathcal{G}$ in our baseline scenario by $\mathcal{G}^{\mathrm{s}}$ and $\mathcal{G}^{\mathrm{m}}$, respectively, we obtain two scenario variants for Scenario 1. The patient and physician populations used in each scenario variant are summarized in Table 5.7.

The simulation results for Scenario 1 in Table 5.8 show a severe deterioration of all patient and physician performance indicators compared to the baseline scenario. The physicians' expected workload measured through the average number of treatments increases by $23\,\%$ for the short-term and $48\,\%$ for the medium-term shift. Due to the increased scarcity of appointments, more and more patients are forced to visit physicians as walk-ins ($56\,\%$ walk-in treatments for short-term and $62\,\%$ for medium-term shift). The average daily overtime for physicians (that neglects all the physicians' administrative and organizational tasks) increases by 2.09 minutes for the short-term and 9.23 minutes for the medium-term shift. On average, patients wait $29\,\%$ longer for their appointments in the short-term and even $66\,\%$ longer in the medium-term shift scenario variant. Similar increases can be observed for the patients' average access distance, which increases by $35\,\%$ to $6.66\,\mathrm{km}$ for the short-term and by $51\,\%$ to $7.51\,\mathrm{km}$ for the medium-term shift. The average waiting time for patients with appointment is almost unaffected by the decline in the number of physicians, which can be explained by the strict prioritization in PFCFS. The average waiting time for walk-ins increases by $30\,\%$ for the short-term and $65\,\%$ for the medium-term shift.

Tab. 5.8.: Mean performance indicators and 95 %-confidence intervals obtained by repeating each simulation experiment 20 times for both variants of Scenario 1.

	Decline in PCPs Short-term Shift	
	Mean	95 %-CI
average number of treatments	12,412.3	[12,399.87, 12,424.73]
average number of walk-ins	6,935.42	[6,923.02, 6,947.81]
average number of acute appointments	2,766.77	[2,762.78, 2,770.77]
average number of regular appointments	2,710.11	[2,705.98, 2,714.25]
average utilization [%]	80.72	[80.65, 80.79]
average daily overtime [min]	2.89	[2.75, 3.02]
average number of rejected walk-ins	69.6	[67.21, 72.92]
average access time [d]	3.18	[3.16, 3.2]
average access time regular [d]	1.6	[1.56, 1.64]
average access distance [km]	6.66	[6.65, 6.66]
average waiting time appt. [min]	2.22	[2.2, 2.23]
average waiting time walk-in [min]	51.51	[51.36, 51.65]
on-time appointments [%]	58.94	[58.86, 59.03]
number of acute illnesses	136,517.25	[136,334.27, 136,700.23]
number of chronic patients	10,662	–
total PCP capacity [h]	26,455	–

	Decline in PCPs Medium-term Shift	
	Mean	95 %-CI
average number of treatments	15,006.28	[14,992.61, 15,019.95]
average number of walk-ins	9,361.44	[9,347.77, 9,375.11]
average number of acute appointments	2,331.72	[2,325.66, 2,337.79]
average number of regular appointments	3,313.12	[3,306.98, 3,319.25]
average utilization [%]	88.51	[88.44, 88.59]
average daily overtime [min]	10.03	[9.79, 10.27]
average number of rejected walk-ins	357.1	[346.92, 368.03]
average access time [d]	4.09	[4.05, 4.12]
average access time regular [d]	1.84	[1.78, 1.89]
average access distance [km]	7.51	[7.5, 7.52]
average waiting time appt. [min]	2.18	[2.16, 2.2]
average waiting time walk-in [min]	65.76	[65.56, 65.96]
on-time appointments [%]	58.54	[58.42, 58.66]
number of acute illnesses	136,499.55	[136,348.12, 136,650.97]
number of chronic patients	10,662	–
total PCP capacity [h]	22,139	–

Tab. 5.9.: Age class distributions for aged patient population.

	16-24	25-65	>65
short-term shift	0.1051	0.6283	0.2666
medium-term shift	0.1025	0.6033	0.2942

5.4 Scenario 2: Aging Patients

Scenario 2 models the ongoing aging of the patient population for a short- and medium-term shift in time. For this purpose, we adjust the discrete probability distribution determining the patients' age classes to generate two new patient populations. More precisely, we use current population projections (Information und Technik Nordrhein-Westfalen, 2019) for the years 2025 and 2030 to obtain the two adjusted discrete probability distributions for the patients' age classes shown in Table 5.9. Using these distributions, we generate the aged patient population $\mathcal{P}^{\mathrm{s}}$ for the short-term and $\mathcal{P}^{\mathrm{m}}$ for the medium-term shift. By replacing the patient population $\mathcal{P}$ in our baseline scenario by $\mathcal{P}^{\mathrm{s}}$ and $\mathcal{P}^{\mathrm{m}}$, respectively, we obtain two scenario variants for Scenario 2; compare Table 5.7.

The simulation results for Scenario 2 (Table 5.10) paint a similar picture as in Scenario 1, i.e., the majority of patient and physician indicators deteriorate, albeit far less severe. As a result of the aging of the patient population, the average number of treatments per physician increases by $1\,\%$ and $2\,\%$ for the short-term and medium-term shift, respectively. However, in contrast to Scenario 1, additional treatments distribute more evenly between appointments and walk-in visits and thus the expected ratio of walk-in treatments to all treatments remains almost unchanged ($47\,\%$ for short-term and $48\,\%$ for medium-term shift). Judging from the almost unaffected average overtime, physicians manage to accommodate the additional treatments mostly within their regular opening hours. As a result of the increased treatment demand, patients wait on average $2\,\%$ longer for their appointments in the short-term and $5\,\%$ longer in the medium-term shift scenario variant. Moreover, they are willing to accept $1\,\%$ ($2\,\%$) longer average access distances in the short-term (medium-term) shift scenario to receive more timely treatment or avoid longer waiting times. Patient waiting times with appointment are unaffected by the increased patient demand. The average waiting times of walk-ins in the short-term shift scenario variant remain almost unchanged, while they increase by $1\,\%$ for the medium-term shift.

5.5 Scenario 3: Combined Effects

Scenario 3 models a combined decline in the number of PCPs and aging of the patient population for a short- and medium-term shift in time. By replacing both, the patient and the physician population in our baseline scenario with the adjusted patient and physician populations from Scenarios 1 and 2, we obtain two scenario variants for Scenario 3; compare Table 5.7.

Tab. 5.10.: Mean performance indicators and 95 %-confidence intervals obtained by repeating each simulation experiment 20 times for both variants of Scenario 2.

	Aging Patients Short-term Shift	
	Mean	95 %-CI
average number of treatments	10,222.16	[10,211.54, 10,232.78]
average number of walk-ins	4,831.4	[4,820.79, 4,842.02]
average number of acute appointments	3,194	[3,192.42, 3,195.58]
average number of regular appointments	2,196.75	[2,195.35, 2,198.16]
average utilization [%]	72.6	[72.54, 72.65]
average daily overtime [min]	0.77	[0.73, 0.81]
average number of rejected walk-ins	14.5	[13.95, 16.18]
average access time [d]	2.52	[2.51, 2.53]
average access time regular [d]	1.51	[1.48, 1.53]
average access distance [km]	5	[5.0, 5.01]
average waiting time appt. [min]	2.11	[2.1, 2.12]
average waiting time walk-in [min]	39.9	[39.74, 40.05]
on-time appointments [%]	60.94	[60.86, 61.01]
number of acute illnesses	137,863.35	[137,692.8, 138,033.9]
number of chronic patients	10,776	–
total PCP capacity [h]	32,617	–

	Aging Patients Medium-term Shift	
	Mean	95 %-CI
average number of treatments	10,300.37	[10,288.82, 10,311.91]
average number of walk-ins	4,909.14	[4,897.62, 4,920.66]
average number of acute appointments	3,162.45	[3,160.73, 3,164.17]
average number of regular appointments	2,228.78	[2,226.99, 2,230.56]
average utilization [%]	72.95	[72.88, 73.03]
average daily overtime [min]	0.8	[0.74, 0.86]
average number of rejected walk-ins	16.4	[15.89, 17.63]
average access time [d]	2.58	[2.57, 2.58]
average access time regular [d]	1.51	[1.49, 1.54]
average access distance [km]	5.04	[5.04, 5.05]
average waiting time appt. [min]	2.11	[2.1, 2.12]
average waiting time walk-in [min]	40.11	[39.97, 40.24]
on-time appointments [%]	60.85	[60.76, 60.93]
number of acute illnesses	138,698.8	[138,516.55, 138,881.06]
number of chronic patients	10,931	–
total PCP capacity [h]	32,617	–

Tab. 5.11.: Mean performance indicators and 95 %-confidence intervals obtained by repeating each simulation experiment 20 times for both variants of Scenario 3.

	Combined Effects Short-term Shift	
	Mean	95 %-CI
average number of treatments	12,536.05	[12,522.14, 12,549.95]
average number of walk-ins	7,059.09	[7,045.14, 7,073.04]
average number of acute appointments	2,743.98	[2,739.24, 2,748.71]
average number of regular appointments	2,732.98	[2,728.25, 2,737.71]
average utilization [%]	81.11	[81.03, 81.2]
average daily overtime [min]	2.94	[2.81, 3.08]
average number of rejected walk-ins	75.15	[72.01, 79.12]
average access time [d]	3.3	[3.27, 3.32]
average access time regular [d]	1.68	[1.63, 1.73]
average access distance [km]	6.74	[6.73, 6.75]
average waiting time appt. [min]	2.21	[2.19, 2.22]
average waiting time walk-in [min]	52.07	[51.91, 52.23]
on-time appointments [%]	58.98	[58.89, 59.07]
number of acute illnesses	137,830.15	[137,657.8, 138,002.52]
number of chronic patients	10,776	–
total PCP capacity [h]	26,455	–

	Combined Effects Medium-term Shift	
	Mean	95 %-CI
average number of treatments	15,269.52	[15,258.29, 15,280.75]
average number of walk-ins	9,624.34	[9,613.2, 9,635.49]
average number of acute appointments	2,257.32	[2,248.85, 2,265.78]
average number of regular appointments	3,387.86	[3,379.37, 3,396.35]
average utilization [%]	89.35	[89.27, 89.43]
average daily overtime [min]	11.32	[11.08, 11.57]
average number of rejected walk-ins	428.9	[421.41, 437.2]
average access time [d]	4.34	[4.3, 4.38]
average access time regular [d]	1.94	[1.88, 2.01]
average access distance [km]	7.54	[7.53, 7.54]
average waiting time appt. [min]	2.15	[2.14, 2.16]
average waiting time walk-in [min]	67.2	[67.01, 67.39]
on-time appointments [%]	58.68	[58.6, 58.77]
number of acute illnesses	138,667.85	[138,534.48, 138,801.22]
number of chronic patients	10,931	–
total PCP capacity [h]	22,139	–

Analyzing our simulation results in Table 5.11 for Scenario 3, we can confirm that the combined effects of a decline in the number of PCPs and an aging population lead to the greatest deterioration of patient and physician indicators among all scenarios. However, the effect of the combined changes compared to the combination of the individual effects from Scenarios 1 and 2 varies between indicators. For the average number of treatments and the ratio of walk-ins, the effects of the combined changes correspond to the sum of the effects for the individual changes, e.g., a $24\,\%$ increase in the average number of treatments in short-term shift variant of Scenario 3 versus a $23\,\%$ and $1\,\%$ increase in the respective variants of Scenarios 1 and 2. Concerning the physicians' average overtime, we can observe that a combined consideration of both changes has an amplifying effect. For example, in the medium-term shift variants of Scenarios 1 and 2 the average overtime increases by 9.23 and 0 minutes, respectively, while the combined changes in Scenario 3 lead to an increase of 10.52 minutes. Similar amplifying effects can be observed for the patients' average access time and walk-in waiting time. Considering the patients' average access distance, the combination of both changes leads to different effects in the two scenario variants. In the short-term shift variant, the effect of the combined changes corresponds to the sum of the effects for the individual changes. In the medium-term shift variant, we can observe a slight dampening effect resulting from a combined consideration of both changes, i.e., while the individual changes lead to a respective $52\,\%$ and $2\,\%$ increase of the expected average access distance, the combined effects lead to an increase of $52\,\%$.

5.6 Sensitivity Analysis

SiM-Care is a complex model that requires a large number of input parameters. While this makes the model very versatile, it simultaneously poses the risk of instabilities and high sensitivities towards small changes in the input values. To ensure that such undesired behaviors do not invalidate the outputs of our simulation experiments, we present a sensitivity analysis. This analysis intends to quantify the changes in the performance indicators resulting from a perturbation of the input parameters.

For the sensitivity analysis of SiM-Care, we consider the baseline scenario in the setup of the case study, i.e., 20 independent runs modeling one year preceded by a warm-up of 60 years. As part of this thesis, we refrain from performing a full sensitivity analysis on all input parameters, but demonstrate the process for those input parameters from Table 5.1 that are least strongly anchored in empirical data. Specifically, we study the model's sensitivity towards the patients' health condition $c \in [0, 1]$ and the age classes' deviation from the illness duration $\Delta_a^d \geq 0$, deviation from the willingness to wait $\Delta_a^\omega \geq 0$, and probability to cancel an appointment after recovery $p_a \in [0, 1]$.

We vary each input parameter relative to its original value in the baseline scenario between $\pm 20\,\%$ in increments of $1\,\%$. To quantify the model's sensitivity, we analyze the resulting impact on the PCPs' average overtime, utilization, and number of rejected walk-ins as well

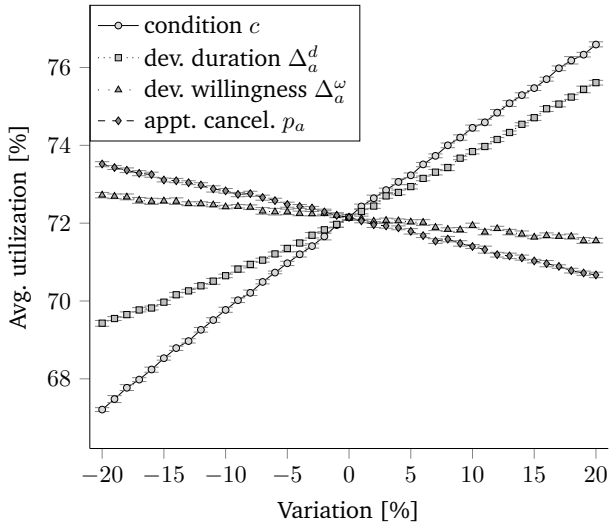


Fig. 5.3.: Mean average utilization and corresponding $95\,\%$ exact confidence intervals for varying input parameters.

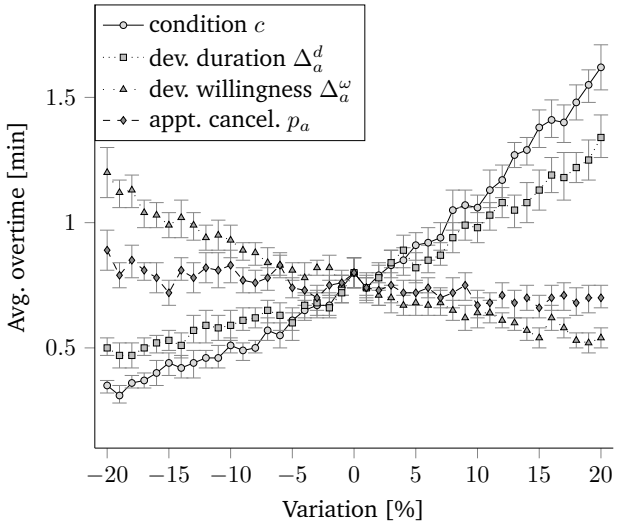


Fig. 5.4.: Mean average overtime and corresponding $95\,\%$ exact confidence intervals for varying input parameters.

as the patients' average access time, access distance, and waiting time with and without an appointment; compare Section 4.5. Figures 5.3–5.9 show the resulting mean values and associated exact $95\,\%$ confidence intervals for all considered performance indicators.

These results do not indicate that the complexity of SiM-Care leads to instabilities. Instead, all performance indicators behave as expected towards variations of the input parameters, e.g., increasing the patients' deviation from the willingness to wait $\Delta_a^\omega \geq 0$ causes patients to wait longer for appointments, which in turn leads to fewer treatments and thus decreases the PCPs' average utilization and overtime.

In terms of the shape of the relationship between changes in input and output, we observe different phenomena. For the average utilization (Figure 5.3), average access time (Figure 5.6), and average waiting time without appointment (Figure 5.9), the relationships is almost linear with rather small confidence intervals. Moreover, the slopes within these relationships are relatively flat, which opposes a high sensitivity towards small changes in the input values.

For the average overtime (Figure 5.4) and the average number of rejected walk-ins (Figure 5.5), the relationship is non-linear with large confidence intervals. The sensitivity and the

Fig. 5.5.: Mean average number of rejected walk-ins and corresponding $95\,\%$ exact confidence intervals for varying input parameters.

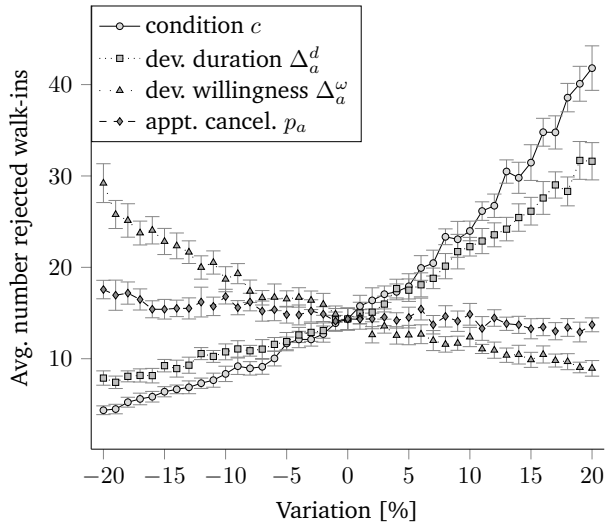


Fig. 5.6.: Mean average access time and corresponding $95\,\%$ exact confidence intervals for varying input parameters.

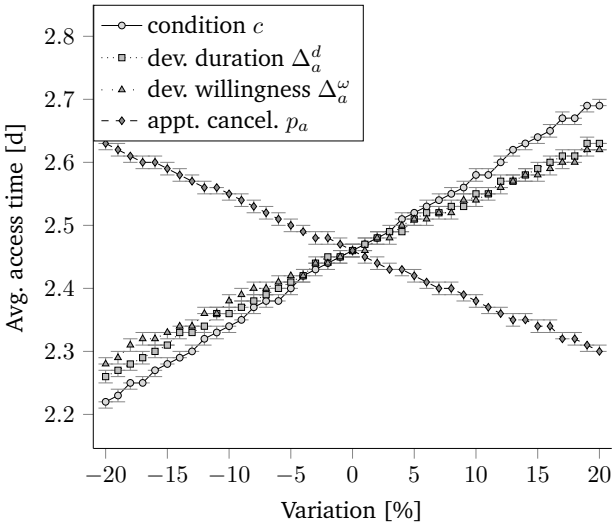
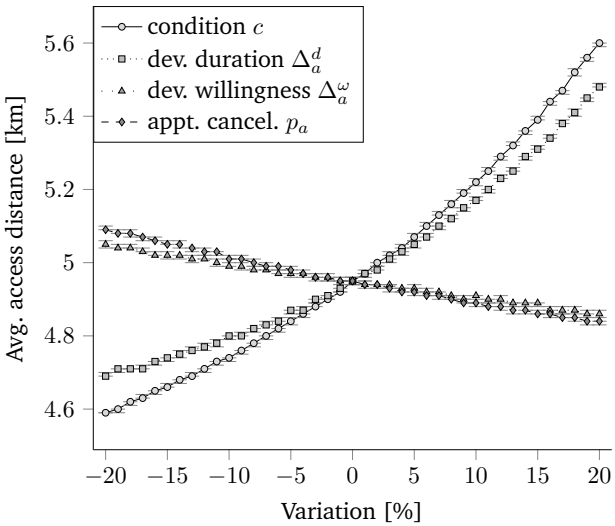


Fig. 5.7.: Mean average access distance and corresponding $95\,\%$ exact confidence intervals for varying input parameters.



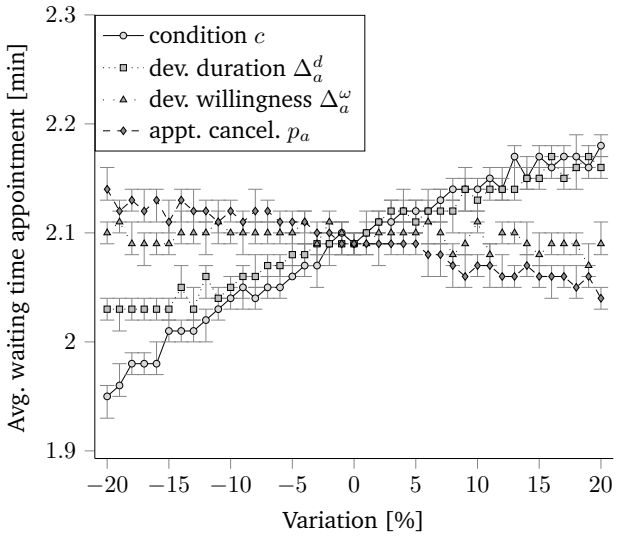


Fig. 5.8.: Mean average waiting time with appointment and corresponding 95 % exact confidence intervals for varying input parameters.

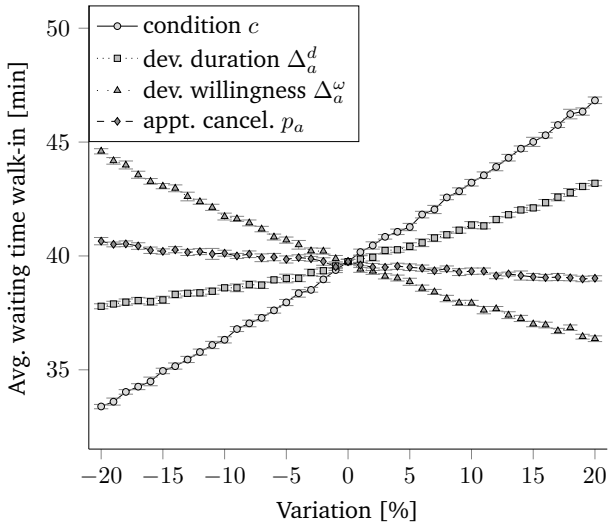


Fig. 5.9.: Mean average waiting time as walk-in and corresponding 95 % exact confidence intervals for varying input parameters.

variation in the number of rejected walk-ins furthermore increase with the system's utilization, which seems intuitive. Still, for small changes ($\pm 5\,\%$) of the considered input values, the sensitivities are not extreme given the width of the confidence intervals.

For the average access distance (Figure 5.7) and average waiting time with appointment (Figure 5.8), we can observe a mix of linear and non-linear relationships. Notably, the average waiting time with appointment appears to be very robust towards variations in the input data. This can be explained by the PFCFS treatment strategy implemented by all physicians that strictly prioritizes patients with appointment.

Summing up, there are no indications that SiM-Care suffers from instabilities. For small changes in the input values, the resulting effects on the evaluated performance indicators seem acceptable. However, the results also show that the model's sensitivity depends on the varied input parameter, the evaluated performance indicator, the implemented PCP strategies, and the system's state. Therefore, we recommend to conduct an individual thorough sensitivity analysis for any implemented scenario, following the lines of this example.

6 Discussion and Conclusion

The aim of SiM-Care is to provide decision makers with a tool to analyze and enhance primary care systems. SiM-Care produces meaningful performance indicators that enable a far more detailed assessment of primary care systems compared to the current approaches based on patient-to-physician ratios. Next to a more accurate evaluation of the status quo, SiM-Care can predict and quantify the influence of policy decisions and changes in the system's population, e.g., an aging of the population or a decline in the number of PCPs as illustrated in Chapter 5. Thereby, the model can particularly take several system changes into account at the same time which enables the analysis of combined effects. As all components of a simulation scenario can be easily adjusted, this opens up a broad field of potential applications ranging from physicians' location planning to the evaluation of specific PCP strategies, e.g., in the field of appointment scheduling. Finally, the modular design of SiM-Care perspectively allows for easy model extensions, e.g., to incorporate prospective new supply concepts such as mobile medical units or telemedicine.

The greatest entry requirement to using SiM-Care is the complex and time-consuming task of generating and validating the input scenarios. As SiM-Care models each agent individually, it requires detailed empirical data which has to be obtained from various parties or, even worse, could be unavailable. Moreover, some model components such as the service time distributions are tailored to the German system and thus might have to be adjusted when using SiM-Care to analyze, e.g., a primary care system in the United States. Each of these adjustments, potentially change the model's behavior and thus require a new validation process to ensure that insights derived from SiM-Care are viable for the studied primary care system.

To help overcoming this entry requirement, we exemplified the scenario generation and validation process for a real-world primary care system in Germany. Particularly, we detailed the generation process of all simulation entities and provided available empirical data sources. Although data availability may vary for other primary care systems, this may hint at where the required empirical data can be obtained. We validated our simulation scenario by comparing its output to available empirical data. To show internal validity, we performed 20 independent runs for each simulation experiment and captured the resulting model variability through confidence intervals. However, we need to stress that additional validation should be performed before actual policy decisions are derived from the presented case study. Such validation measures should particularly include an extended sensitivity analysis as well as an expert validation which were out of scope for the purpose of this thesis.

In summary, SiM-Care can serve as a versatile decision support tool in primary care planning when it is used with adequately validated simulation scenarios. The process of generating and validating simulation scenarios is both challenging and time-consuming. However, once this process is complete, SiM-Care enables a detailed analysis and evaluation of primary care systems that is superior to basic ratio-based approaches. As a final motivation to use SiM-Care, we present a potential real-world use-case from Germany. At the beginning of 2019, the German Bundestag passed the law for faster appointments and better care (TSVG) that increased the minimum weekly opening hours for physicians with statutory health insurance accreditation from 20 to 25 hours (Bundestag, 2019, Art. 15). The law is controversial, among other things, because there are doubts about the consequences of this policy decision (Korzilius, 2019). Using SiM-Care, decision makers could have obtained insights into the effects of increased minimum opening hours from both the patients' and the physicians' perspective before its implementation.

Future work should include further efforts towards model validation and calibration as well as the implementation of model extensions. Currently, illness distributions are considered static by SiM-Care. By modeling dynamic illness distributions, we can make them dependent on seasonality or the patients' previous history of illnesses. In the current model, the duration of an illness is independent of the actual treatment. Interestingly, the results are convincing even without this causal link. In the future, we want to compare whether implementing this link in the conceptual model significantly affects our findings. A similar comparison shall investigate the influence of no-show patients, who introduce unexpected idle time into the physicians' schedules. The sole integration of patient non-attendance into SiM-Care is straightforward, as it suffices to generate arrival events stochastically. The actual difficulty lies in the need for empirical no-show probabilities as well as the necessity to decide how no-show patients continue their course of treatment and how PCPs anticipate non-attendance through more sophisticated strategies. Yet other possible model extensions include: illness specific appointments such that not all acute illnesses are treated during every appointment, intentional physicians' breaks, implementation of additional patient attributes such as gender, and mobile patient agents that move between different locations, e.g., their home and work. Finally, we are currently preparing the open source release of our model implementation that comes with a graphical user interface (Figure 6.1) such that SiM-Care can be easily accessed, studied, and adapted to the individual requirements of all modelers.

Since its first publication, SiM-Care has already been extended and applied to evaluate the use of information and communication technologies (ICT) in primary care (Büsing et al., 2020b). To that end, the authors extended patient agents by an attribute that measures the patient's affinity towards ICT which was collected within the framework of a nationwide trend study. Moreover, they developed two new appointment systems that support digital ICT. The experiments conducted in SiM-Care revealed that senior citizens with their currently estimated ICT usage behavior are likely to be placed at a disadvantage when digital appointment scheduling systems are introduced.

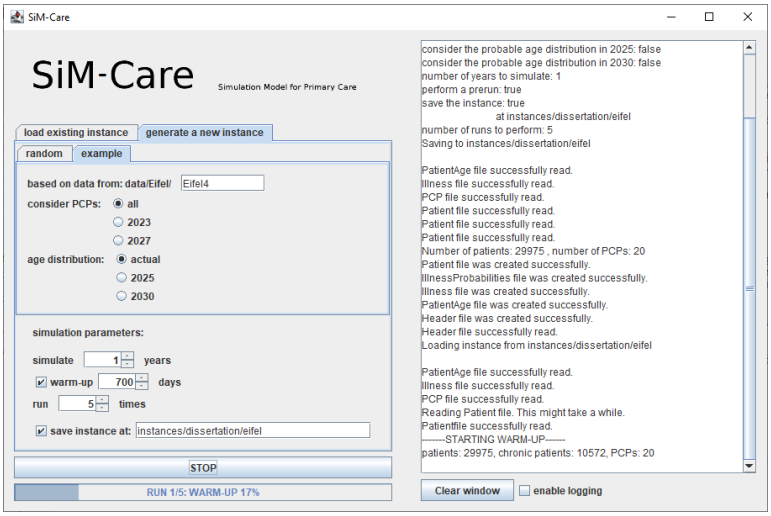


Fig. 6.1.: Graphical user interface (GUI) of SiM-Care.

As part of an ongoing work on optimized appointment systems, Büsing et al. (2020a) implemented and evaluated so-called mask-based appointment systems in SiM-Care which preallocate appointment slots to certain types of patient requests. Furthermore, the authors integrated robust mask-based appointment systems into SiM-Care which can adapt to fluctuations in patient demands by switching between preoptimized scheduling templates.

Last but not least, an extension of SiM-Care that can be used to evaluate the effects of mobile medical units on primary care systems is introduced in the subsequent Part II of this thesis.

Part II

Operational Planning for Mobile Medical Units

A Robust Three-Phased Optimization Approach

7 Introductory Remarks and Contribution

> *The fundamental problem of rural health is the extremely low density of health professionals per unit of surface area, resulting in large distances between patient and services.*
>
> **— Thomas S. Bodenheimer, 1969**

7.1 Motivation and Research Question

The geographic-demographic truths of rural health have always manifested themselves in a low density of health professionals and thus long access distances for patients (Bodenheimer, 1969). However, as the number of physicians continues to decline while the needs of the aging populations increase, existing barriers to health services are at risk of multiplying manifold (Mann et al., 2010; Alemayehu and Warner, 2004).

To counteract the growing distances between patients and services, the overcoming of access barriers with the help of mobile medical units (MMUs) has been studied in many developed and developing countries; compare Bodenheimer (1969), Thorsen and McGarvey (2018), Khanna and Narula (2017), Hill et al. (2014), and Schwartze and Wolf (2017). MMUs (also known as mobile health facilities or mobile clinics) are customized vehicles fitted with medical equipment that are easy to relocate and that can provide most of the health services a regular stationary practice could; see Figure 7.1. The flexibility of MMUs offers the possibility of a local and demand-oriented provision of primary care in sparsely populated regions, which are characteristic to rural settings (Hill et al., 2014).

Although MMUs have been reported to operate in various modes, we will focus exclusively on a weekly recurring operation in clinical sessions as described in Schwartze and Wolf (2017), Bodenheimer (1969), and Thorsen and McGarvey (2018). In this mode of operation, MMUs are stationed in larger cities (which they do not serve) and set out each day to provide health services at fixed sites in the surrounding rural communities. As common in primary care, we thereby structure each day into a morning and an afternoon session (Klassen and Rohleder, 1996) and thus MMUs can service at most two sites per day. At the end of each day, MMUs return to their home depot such that all personnel can return to their homes overnight. Next to the benefit of increased staff satisfaction, this incidentally reduces the cost for nonexempt

Fig. 7.1.: Mobile medical unit operated in rural parts of India.[3]

staff (Thorsen and McGarvey, 2018). While there are different types of MMUs, e.g., for off-road operations, we restrict our considerations to a fleet of identical vehicles.

The potentials of MMUs are demonstrated by the increasing number of applications in practice. The US alone has an estimated $1{,}500$ MMUs receiving 5 million visits each year (Hill et al., 2014). Nevertheless, we must not forget that MMUs are still nascent to health care delivery and that their operation is often associated with major challenges (Khanna and Narula, 2017). For instance, Patro et al. (2008) report long patient waiting times as a result of a high workload while other studies faced problems with small or decreasing number of patient visits (Schwartze and Wolf, 2017; Geoffroy et al., 2014). There is thus a general consensus, that better strategies for the prelaunch of an MMU service are required (Khanna and Narula, 2017). In a logical consequence we formulate the following question:

> How can the effectiveness and sustainability of an MMU service be improved by optimized prelaunch strategies?

Developing prelaunch strategies for an MMU service is highly non-trivial, as the flexibility of MMUs necessitates a complex planning problem: Operation sites for MMUs have to be set up to provide essential external infrastructure such as waiting rooms; weekly recurring MMU sessions must be scheduled to meet uncertain patient demands; and daily vehicle routes covering these sessions must be planned. All these planning decisions are coupled and jointly determine the operation costs which need to be minimized to ensure a sustainable service. In the following, we present a prelaunch strategy for an MMU service that accounts for demand uncertainties and consists of three sequential planning phases.

MMU Operational Planning in Three Phases

This thesis introduces the integrated solution framework P3MMU for the operational planning of MMUs that decouples planning decisions into three sequential phases. The solution method thereby trades-off potentially suboptimal operation cost in exchange for simpler subproblems that can be solved by tailored algorithms to increase the framework's overall computational

[3] Wikimedia Commons under CC0.

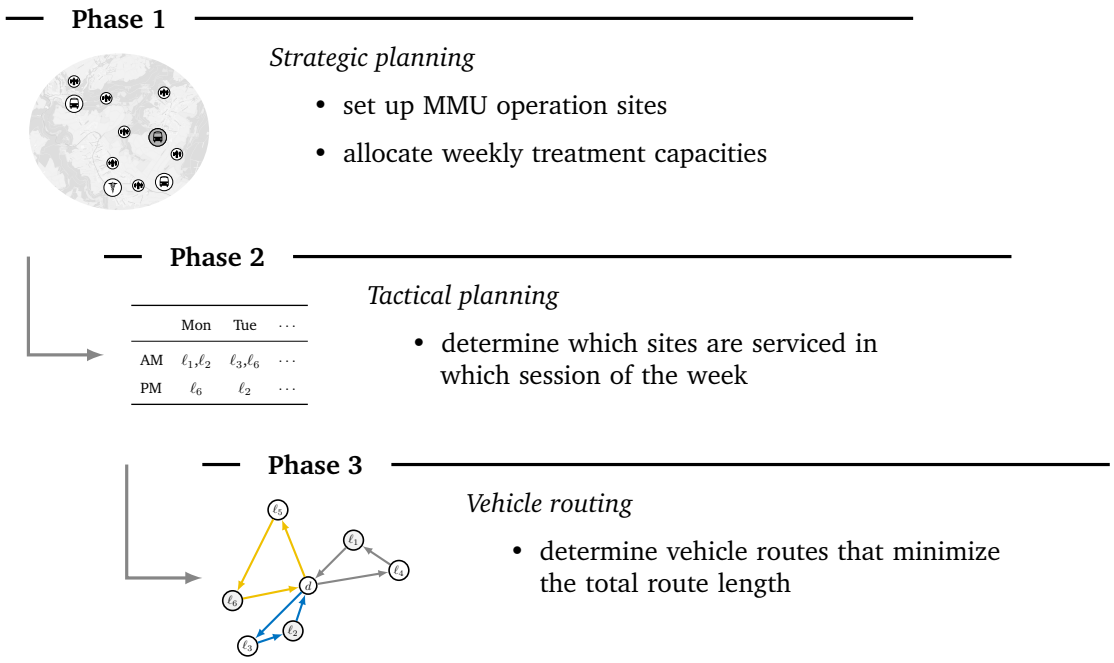


Fig. 7.2.: Phases in P3MMU for the operational planning of mobile medical units.[4]

performance. In the following, we introduce and describe each phase individually. A graphical overview over the three sequential phases in P3MMU can be found in Figure 7.2.

Phase 1 considers the strategic problem of deciding where MMU operation sites should be set up and how often these sites should be serviced in the course of a week. As potential MMU operation sites are generally pre-allocated, e.g., by the serviced municipalities (Schwartze and Wolf, 2017), the set up sites are chosen from a given set of candidates. The number of sessions that can be operated per week at each site is limited and the allocation of treatment capacities must consider the existing practices as well as the uncertain patient demands.

Phase 2 addresses the tactical problem of determining the serviced MMU operation sites for each session of the week by partitioning the weekly MMU operations which are determined in Phase 1. This partitioning must ensure an even distribution throughout the week such that the number of required vehicles as well as the patients' maximum daily access distances are minimized which leads to a bottleneck optimization problem.

Phase 3 deals with the problem of determining the vehicle routes for each day of the week that serve the partitioned MMU sessions as they were obtained from Phase 2. As previously established, every MMU may operate at most two sessions per day (one in the morning and one in the afternoon) and all routes must start and end in the same depot. The total length of all routes must be minimized to reduce operation cost.

[4]Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

Phases 1, 2, and 3 interconnect in P3MMU and are solved sequentially, i.e., Phase 1 provides the input for Phase 2 which in turn provides the input for Phase 3. The following section summarizes our contribution and outlines our solution approaches before we provide an overview of the related literature.

7.2 Contribution

The main contribution of Part II of this thesis is the optimization framework P3MMU for the operational planning of MMUs. The approach consists of three sequential phases that interconnect and structure MMU operational planning into three subproblems. We provide solution approaches for each of these three subproblems and evaluate P3MMU in a computational study for a set of test instances based on a rural real-world primary care system in Germany. The study of the three subproblems yields multiple subcontributions that we summarize hereinafter.

In Phase 1, we investigate the *strategic planning problem for MMUs* (SMMU) – a capacitated set covering problem that includes existing practices and two types of patient demands: i) steerable demands representing patients who seek health services through a centralized appointment system and can be steered to any treatment facility within a given consideration set and ii) unsteerable demands representing walk-ins who always visit the closest available treatment facility. We present a compact linear integer programming formulation for the SMMU which we subsequently solve via Benders decomposition and constraint generation. To account for uncertainties in both types of patient demands, we introduce exact constraint generation algorithms to solve the robust counterpart of the Benders formulation for interval and budgeted uncertainty sets. To the best of our knowledge, this is the first contribution that studies the allocation of MMUs at a strategic level as a robust set covering problem. The concept of modeling two different types of demands has not been considered so far and represents a new extension to the field of location planning.

As part of Phase 2, we study a bottleneck location problem called the *tactical partitioning problem for MMUs* (TPMMU) which is a partitioning variant of the k-center problem. Using a reduction from the dominating set problem, we show that the TPMMU is $\mathcal{NP}$-hard to approximate within a constant approximation factor. Moreover, we show how an adaptation of this reduction implies that the metric problem variant of the TPMMU is $\mathcal{NP}$-hard to approximate within a constant approximation factor $1 < \alpha < 2$. As a solution approach, we present a compact mixed-integer linear programming formulation of the TPMMU. Alternatively, we show how Phases 1 and 2 can be jointly solved by extending our solution approaches for the SMMU to a session-specific problem variant to which we refer as the *combined strategic tactical planning problem for MMUs* (STMMU).

Finally, we consider the *vehicle routing problem for MMUs with a single depot* (VRMMU) and the *vehicle routing problem for MMUs with multiple depots* (mVRMMU) in Phase 3. In the

single-depot setting, we reduce the VRMMU to a minimum weight perfect matching problem in a bipartite gadget which can be solved in polynomial time. For the multi-depot setting, we show that the mVRMMU is special case of the so-called *budgeted colored bipartite perfect matching problem* (BCBPM) which we subsequently prove to be strongly $\mathcal{NP}$-hard. To solve the mVRMMU, we derive a compact binary linear programming formulation.

7.3 Related Work

To the best of our knowledge, one of the earliest works on the operational planning of MMUs is due to Hodgson et al. (1998) who consider a location-routing problem for a single MMU, i.e., the authors simultaneously determine the vehicle stops and the vehicle route. In a feasible solution, a set of population centers must be within a prespecified distance of a vehicle stop along the planned route and the objective is to minimize the total length of the route. A binary programming formulation of the problem is presented and subsequently solved exactly by a branch-and-cut procedure or heuristically using a generalized insertion algorithm. In a follow-up article, Hachicha et al. (2000) extend this problem setting to multiple vehicles and vehicle stops that must be serviced. All MMU routes must start and end at a central depot and the number of stops per route as well as the total route length is bounded to ensure a balanced workload between MMUs. Doerner et al. (2007) consider the problem formulation in Hodgson et al. (1998) for multiple objective functions. That is, they evaluate MMU routes with respect to three criteria: economic efficiency, average access distance, and coverage. The corresponding Pareto fronts are approximated using ant colony optimization and genetic algorithms. Moreover, the authors also consider multiple MMUs, however only by dividing the population centers into multiple service areas which are then each serviced by a separate vehicle. Ozbaygin et al. (2016) extend the coverage objective for the one vehicle setting to partial coverage, i.e., only the population centers that are visited by an MMU are completely covered while population centers in reach of an MMU stop are only covered with a certain percentage. More recently, Yücel et al. (2018) further generalize the idea of partial coverage to multiple vehicles and integrate their MIP formulation into a data-driven optimization framework based on credit card transactional data.

The main difference between the previous articles and this thesis is the considered setting: While the former focus on very extensive regions with bad road infrastructure and MMU routes that can be multiple weeks long, we consider the problem on a much smaller scale with MMU routes that service at most two stops per day and return to a depot each night. As a result, the vehicle routing plays a far more important role in Hodgson et al. (1998), Hachicha et al. (2000), Doerner et al. (2007), Ozbaygin et al. (2016), and Yücel et al. (2018) and is therefore considered at the strategic level while the incorporation of demands and allocation of treatment capacities are considered downstream once the MMU routes are fixed. We, on the other hand, consider patient demands and the allocation of treatment

capacities at the strategic level and shift the vehicle routing into Phase 3 which boils down to a polynomial-time solvable matching problem in the single-depot setting.

A problem originating in humanitarian logistics that is quite related to the operational planning of MMUs is studied by Tricoire et al. (2012). In this problem, the authors consider the simultaneous setup of distribution centers and the routing of vehicles that restock these with relief goods. Distribution centers have a certain capacity and their setup induces costs. The demand for relief goods at the population centers is uncertain and targets the closest distribution center. The objective is the minimization of the setup and routing costs while the expected coverage of the demands is to be maximized. The problem is formulated as a stochastic bi-objective combinatorial optimization problem and solved by combining a scenario-based approach with an epsilon-constraint method. A deterministic single objective variant of this problem that does not consider setup costs for distribution centers and allows demands to be freely assigned among all operated distribution centers within a certain covering distance is studied by Naji-Azimi et al. (2012). Instead of maximizing the coverage, full-coverage is enforced while the total routing cost is minimized.

Comparing our problem setting to the problems in Tricoire et al. (2012) and Naji-Azimi et al. (2012), we note that neither of the latter consider existing infrastructure and only either unsteerable or steerable demands but not the combination of the two. Moreover, both problems put a strong emphasis on the vehicle routing, which is less important in our setting and thus considered in the last phase of P3MMU.

The SMMU studied in Phase 1 is a pure covering location problem, more specifically a set covering problem. Set covering problems have been studied extensively in various applications and comprehensive review articles on existing work in this field can for example be found in Caprara et al. (2000), Farahani et al. (2012), Ahmadi-Javid et al. (2017), and García and Marín (2015). In the following, we will focus our review on set covering problems that incorporate uncertainties in a robust or probabilistic optimization framework.

Probably the most related set covering problem to the SMMU is the q-multiset multicover problem studied by Krumke et al. (2019). The q-multiset multicover problem is the special case of the decision version of the SMMU that does not consider existing facilities, setup cost, and unsteerable demands. The authors study the problem's complexity and investigate the problem's extension to uncertain demands that may vary within a given interval. Using budgeted uncertainty sets, they devise a formulation of the robust counterpart which can be solved by constraint generation. This thesis builds on the results in Krumke et al. (2019) and generalizes them to the SMMU. Various other studies on robust set covering problems mostly differ in terms of the applied robustness concept. Dhamdhere et al. (2005) introduce demand-robust covering problems and provide approximation algorithms. Feige et al. (2007) consider two-stage robust covering problems and devise approximation algorithms, whereas Gupta et al. (2014) study approximation algorithms for the k-robust set covering problem. The set covering problem with uncertain cost coefficients is considered by Pereira and Averbakh (2013) and exact algorithms for computing min-max regret solutions are presented.

Right-hand side uncertainties in set covering problems in the form of chance constraints, i.e., where demands only have to be covered with a certain probability, are considered under the name probabilistic set covering problem. Beraldi and Ruszczyński (2002) study the probabilistic set covering problem and devise exact methods by enumerating over the set of p-efficient points. Later, Saxena et al. (2010) introduce the notion of p-inefficiency to devise compact MIP formulations for the probabilistic set covering problem that can be strengthened by separating so-called polarity cuts. Left-hand side uncertainties in set covering problems in the form of chance constraints have been considered under the name uncertain set covering problems. A polyhedral study of the uncertain set covering problems is performed by Fischetti and Monaci (2012) who compare a compact versus a cutting plane model. More recently, Lutter et al. (2017) introduce compact and non-compact robust formulations for the uncertain set covering problem by combining concepts from robust and probabilistic optimization.

For more literature on set covering problems under uncertainty, we refer to the references in Lutter et al. (2017). Preference orderings of clients that are similar to our concept of unsteerable patient demands, have been studied for a deterministic facility location problem known as the simple plant location problem in Hanjoul and Peeters (1987) or more recently in Cánovas et al. (2007).

To the best of our knowledge, there are only two previous articles that consider the allocation of MMUs as a set covering problem. Aguwa et al. (2018) focus on data analytics and reduce the MMU allocation to the standard set covering problem. A more elaborate maximum covering problem for the strategic planning of a single mobile dentistry clinic is considered by Thorsen and McGarvey (2018) with the goal of improving accessibility while maintaining financial sustainability. As both of these works consider purely deterministic settings, this thesis represents the first contribution to the field that considers the strategic allocation of MMUs as a robust set covering problem.

The TPMMU studied in Phase 2 is a bottleneck location problem that can be seen as a partitioning variant of the (discrete) k-center problem. The k-center problem asks to locate k facilities (or centers) that minimize the maximum distance to a set of demand sites and was first introduced by Hakimi (1965). Hsu and Nemhauser (1979) showed that it is $\mathcal{NP}$-hard to approximate the metric k-center problem with a constant approximation factor $1 < \alpha < 2$. This inapproximability result is tight, i.e., there exist various 2-approximation algorithms for the metric k-center problem (Hochbaum and Shmoys, 1985; Mihelič and Robič, 2003). Lim et al. (2005) study the k-center problem with minimum coverage in which centers are required to service a minimum number of clients. As a counterpart, Khuller and Sussmann (2000) study the capacitated k-center problem in which the number of clients that can be served by each center is limited by an upper bound. Comprehensive reviews on the k-center problem can be found for example in Tansel (2011) and Calik et al. (2015).

The fundamental difference between the k-center problem and the TPMMU can be summarized as follows. In the former, we select a subset (of cardinality at most k) of the available centers that minimizes the bottleneck and ignore all unselected centers. In the latter, we must

partition all available centers into a predefined number of groups such that the maximum bottleneck among all groups is minimal. To the best of our knowledge, such partitioning variants of the k-center problem have not yet been considered in the literature.

The vehicle routing problems VRMMU and mVRMMU considered in Phase 3 are structurally specialized matching problems. Related budgeted and colored matching problems are studied in Part III of this thesis. Particularly the multi-budgeted matching problem (mBM) introduced in Chapter 14 generalizes the BCBPM and the presented dynamic programs for the mBM can therefore be used to solve the mVRMMU on special graph classes. For further related work, we refer to the literature review of Part III in Chapter 13.

7.4 Outline and Use of Published Materials

Part II of this thesis is structured as follows. Chapter 8 considers Phase 1, i.e., the strategic planning of MMUs with steerable and unsteerable patient demands. The tactical planning phase that determines the operations in each session (Phase 2) is addressed in Chapter 9. Chapter 10 considers the single- and multi-depot variant of the vehicle routing for MMUs (Phase 3). To evaluate the P3MMU framework in its entirety and demonstrate how the individual phases interconnect, we present a case study in Chapter 11. Finally, we summarize and discuss the results of this part in Chapter 12 and provide directions for future research.

Chapter 8 and parts of Chapters 7, 9, 11, and 12 are based on the publication Büsing et al. (2021) and are therefore joint work with my supervisor Christina Büsing and fellow Ph.D. students Eva Schmidt and Manuel Streicher. Parts of Chapter 10 are based on the publication Büsing and Comis (2018a) which is joint work with my supervisor Christina Büsing.

8 Phase 1: Robust Strategic Planning for MMUs

In this chapter, we consider strategic planning phase in the three-phased optimization framework P3MMU. To that end, we study the combined location and capacity planning for MMU services in form of a capacitated set covering problem called the strategic planning problem for MMUs (SMMU). The SMMU addresses the problem of deciding where MMU operation sites should be set up and how often these should be serviced in the course of a week. As MMUs are intended as a complementary form of health provision that should be integrated into the present primary care systems (Doerner et al., 2007), we include existing practices with their treatment capacities into our model. In addition, we consider two fundamentally different types of patient demands that are common in primary care systems: i) patients who seek health services via a centralized appointment system and can be steered towards any available treatment facility within the patients' consideration sets and ii) walk-ins who forgo the appointment system and always visit the treatment facility of their choice – which we assume to be the closest to the patient. Given the nature of these patients, we refer the former as the *steerable patient demands* while we call the latter the *unsteerable patient demands*.

The main focus of this chapter is the extension of the SMMU to uncertain patient demands which are intrinsic to the nature of health care needs (Fone et al., 2003). To that end, we model both types of patient demands as random variables that we integrate into our models in a robust optimization framework.

The remainder of this chapter is structured as follows. First, we formalize the SMMU in Section 8.1 by providing an integer programming formulation which we solve via Benders decomposition. Subsequently, we extend the problem to uncertainties in both, the steerable and the unsteerable demands in Section 8.2.

8.1 Problem Classification and Formulation

The strategic planning problem for MMUs is a capacitated set covering problem that provides the basis for an MMU service: given a set of potential MMU operation sites L, a set of existing primary care practices P, and a set of aggregated patient demand origins V, decide how many MMU sessions shall be operated at each site $\ell \in L$ in the course of a week in order to meet all patient demands at minimum cost. Potential MMU operation sites $\ell \in L$ have to be set up at cost $c_\ell \in \mathbb{N}$ and allow for up to $b_\ell \in \mathbb{N}$ operated sessions per week. Each operated

MMU session yields a weekly treatment capacity $\hat{b} \in \mathbb{N}$ and induces the cost $\hat{c} \in \mathbb{N}$. Thus, we can define the following.

Definition 8.1. A *strategic MMU operation plan* is a function $m^{\mathrm{S}} \colon L \to \mathbb{N}$ that respects the session capacity at each site, i.e., $m^{\mathrm{S}}_\ell \leq b_\ell$ for all $\ell \in L$ where we notate $m^{\mathrm{S}}_\ell \coloneqq m^{\mathrm{S}}(\ell)$. The *cost* of a strategic MMU operation plan m^{S} is defined by the costs of setting up sites and operating sessions, i.e., $c(m^{\mathrm{S}}) \coloneqq \sum_{\ell \in L : m^{\mathrm{S}}_\ell > 0} c_\ell + \hat{c}\, m^{\mathrm{S}}_\ell$.

Existing primary care practices $p \in P$ have an individual weekly treatment capacity $\bar{b}_p \in \mathbb{N}$. Patient demand origins $v \in V$ specify the weekly treatment demand of a particular region. To prevent patients from having to travel excessive distances, a *consideration set* $N(v) \subseteq L \cup P$ specifies for every demand origin $v \in V$ the feasible treatment facilities. The weekly patient demand at each demand origin $v \in V$ consists of two types of demands: i) the steerable demands $d_v \in \mathbb{N}$ corresponding to patients who announce themselves through a centralized appointment system and can be steered to any operating treatment facility in the consideration set $N(v)$ and ii) the unsteerable demands $u_v \in \mathbb{N}$ corresponding to walk-ins that always visit the nearest operating treatment facility $k^{\min}_v(m^{\mathrm{S}}) \in N(v)$ which depends on a given distance measure $\mathrm{dist} \colon V \times (L \cup P) \to \mathbb{N}$ and the strategic MMU operation plan m^{S}. As part of our model, the nearest operating treatment facility $k^{\min}_v(m^{\mathrm{S}})$ is always be unique and we will see later-on how this is ensure by the definition of an order on $N(v)$.

In a feasible strategic MMU operation plan, all steerable patient demands have to be assigned to a feasible treatment facility and every facility's treatment capacity has to be respected. To formalize these requirements, we first define an assignment of the steerable patient demands to the treatment facilities.

Definition 8.2. An *assignment* of the steerable patient demands is a set of functions $\{f_v\}_{v \in V}$ with $f_v \colon N(v) \to \mathbb{N}$ that distribute all steerable patient demands within their respective consideration set, i.e., $\sum_{k \in N(v)} f_v(k) = d_v$ for all $v \in V$.

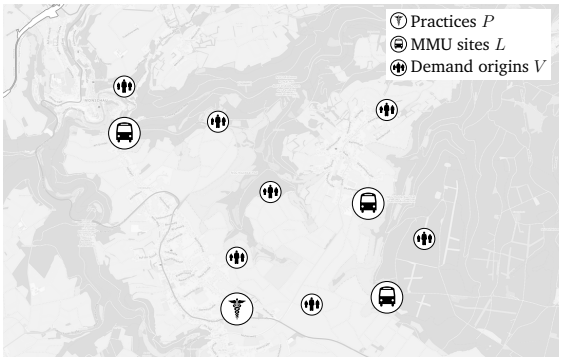
Next, we define *feasible* strategic MMU operation plans. To ease notation, we denote all patient demand origins that can target a treatment facility $k \in L \cup P$ by $N(k) \coloneqq \{v \in V : k \in N(v)\}$.

Definition 8.3. A strategic MMU operation plan m^{S} is *feasible* if there exists an assignment of the steerable patient demands $\{f_v\}_{v \in V}$ that respects the treatment capacity at each treatment facility $k \in L \cup P$; that is

$$\sum_{v \in V : k^{\min}_v(m^{\mathrm{S}}) = k} u_v + \sum_{v \in N(k)} f_v(k) \leq \begin{cases} \bar{b}_k & \text{if } k \in P, \\ \hat{b}\, m^{\mathrm{S}}_k & \text{if } k \in L. \end{cases}$$

Using the notion of a feasible MMU operation plan, we can finally provide a formal definition for the SMMU.

Fig. 8.1.: Domain of the SMMU.[5]



Definition 8.4 (SMMU)**.** Let the potential MMU operation sites $\ell \in L$ with setup costs $c_\ell \in \mathbb{N}$ and weekly session capacities $b_\ell \in \mathbb{N}$ be given. Moreover, let $p \in P$ be the existing practices with weekly treatment capacities $\bar{b}_p \in \mathbb{N}$, and $v \in V$ be the patient demand origins with consideration sets $N(v) \subseteq L \cup P$ and weekly steerable and unsteerable demands $d_v, u_v \in \mathbb{N}$. Then, the *strategic planning problem for MMUs* (SMMU) asks for a feasible strategic MMU operation plan of minimum cost, where every operated MMU session induces the cost $\hat{c} \in \mathbb{N}$ and yields a weekly treatment capacity $\hat{b} \in \mathbb{N}$.

The domain of the SMMU is illustrated in Figure 8.1. We begin the classification of the SMMU, by showing that the problem is strongly $\mathcal{NP}$-hard.

Theorem 8.5. *The SMMU is strongly $\mathcal{NP}$-hard.*

Proof. By setting $\hat{b} = 3$, $P = \emptyset$, $\hat{c} = 1$, $u_v = 0$ for all $v \in V$, $c_\ell = 0$ for all $\ell \in L$ and choosing b_ℓ large enough, e.g., $b_\ell = \sum_{v \in V} d_v$, it becomes evident that 3-multiset multicover is a special case of the decision version of the SMMU. Thus, the strong $\mathcal{NP}$-hardness result for the SMMU follows directly from the corresponding result for 3-multiset multicover in Krumke et al. (2019). □

To solve the SMMU, we present an integer linear programming formulation that we subsequently solve by a Benders decomposition approach. Let variables $y_\ell \in \{0, 1\}$ indicate whether site $\ell \in L$ is set up, let variables $x_\ell \in \mathbb{N}$ denote the number of weekly operated MMU sessions at site $\ell \in L$, and let variables $z_{vk} \in \mathbb{N}$ determine the weekly steerable demand originating in demand origin $v \in V$ that is assigned to the treatment facility $k \in N(v)$. Moreover, let variables $w_{vk} \in \{0, 1\}$ indicate the closest operating treatment facility $k \in N(v)$ that is targeted by all unsteerable demands originating in $v \in V$. To that end, let $\pi_v \colon \{1, \ldots, |N(v)|\} \to N(v)$ define an order on the consideration set $N(v)$ that is non-decreasing with respect to the treatment facility's distance $\mathrm{dist} \colon V \times (L \cup P) \to \mathbb{N}$ to demand origin $v \in V$. As a result, $\pi_v(1) \in N(v)$ denotes the closest treatment facility to demand origin $v \in V$ and π_v guarantees the uniqueness of $k_v^{\min}(m^{\mathrm{S}}) \in N(v)$. To ease notation, we denote all potential MMU operation sites and practices within the consideration set of demand origin $v \in V$ by $N_L(v) \coloneqq N(v) \cap L$ and $N_P(v) \coloneqq N(v) \cap P$, respectively. We can now formulate the SMMU as follows:

[5] Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

$$
\begin{aligned}
(\mathrm{Det}) \quad \min_{y,x,z,w} \quad & \sum_{\ell \in L} c_\ell \, y_\ell + \sum_{\ell \in L} \hat{c} \, x_\ell && && (8.1a)\\
\text{s.t.} \quad & x_\ell \leq b_\ell \, y_\ell && \forall \ell \in L && (8.1b)\\
& \sum_{k \in N(v)} z_{vk} \geq d_v && \forall v \in V && (8.1c)\\
& \sum_{v \in N(\ell)} z_{v\ell} + \sum_{v \in N(\ell)} u_v \, w_{v\ell} \leq \hat{b} \, x_\ell && \forall \ell \in L && (8.1d)\\
& \sum_{v \in N(p)} z_{vp} + \sum_{v \in N(p)} u_v \, w_{vp} \leq \bar{b}_p && \forall p \in P && (8.1e)\\
& \sum_{k \in N(v)} w_{vk} \geq 1 && \forall v \in V && (8.1f)\\
& w_{v\ell} \leq y_\ell && \forall v \in V, \, \forall \ell \in N_L(v) && (8.1g)\\
& w_{v\ell} \geq y_\ell - \sum_{i=1}^{\pi_v^{-1}(\ell)-1} w_{v,\pi_v(i)} && \forall v \in V, \, \forall \ell \in N_L(v) && (8.1h)\\
& w_{vp} \geq 1 - \sum_{i=1}^{\pi_v^{-1}(p)-1} w_{v,\pi_v(i)} && \forall v \in V, \, \forall p \in N_P(v) && (8.1i)\\
& x_\ell \in \mathbb{N}, \; y_\ell \in \{0,1\} && \forall \ell \in L && (8.1j)\\
& w_{vk} \in \{0,1\}, \; z_{vk} \in \mathbb{N} && \forall v \in V, \, \forall k \in N(v). && (8.1k)
\end{aligned}
$$

In this formulation, constraints (8.1b) enforce the session capacity at each set up site, inequalities (8.1c) model the assignment of the steerable patient demands, and constraints (8.1d)–(8.1e) guarantee that the treatment capacities at each treatment facility are adhered to. Moreover, inequalities (8.1f)–(8.1i) ensure that unsteerable patient demands target their closest considered operating treatment facility.

We show that the presented integer linear program (Det) is indeed a formulation for the SMMU. To that end, we prove that there always exists an optimal solution (y, x, z, w) to (Det) in which all unsteerable demands originating in $v \in V$ target the closest operated treatment facility $k_v^{\min}(y) \coloneqq \arg\min_{k \in N(v): k \in P \vee (k \in L \wedge y_k = 1)} \mathrm{dist}(v, k)$ in the consideration set.

Lemma 8.6. *Given a feasible solution (y, x, z, w) to (Det), we can compute a feasible solution (y, x, z, w') to (Det) with the same objective value such that for every demand origin $v \in V$ there is at most one treatment facility $k \in N(v)$ with $w'_{vk} = 1$, i.e., $\sum_{k \in N(v)} w'_{vk} \leq 1$, in linear time.*

Proof. Given a feasible solution (y, x, z, w) to (Det), we set

$$
w'_{vk} = \begin{cases} 0 & \text{if} \quad \exists k' \in N(v) : \pi_v^{-1}(k') < \pi_v^{-1}(k) \wedge w_{vk'} = 1, \\ w_{vk} & \text{else.} \end{cases}
$$

Clearly, w' satisfies $\sum_{k \in N(v)} w'_{vk} \leq 1$ and can be computed in linear time. It holds that $w' \leq w$ and thus (y, x, z, w') satisfies inequalities (8.1d) and (8.1e). Moreover, the solution (y, x, z, w') obviously satisfies constraints (8.1f) and (8.1g). Concerning constraints (8.1h), assume there exist $v \in V$ and $k \in N_L(v)$ such that

$$w'_{vk} < y_k - \sum_{i=1}^{\pi_v^{-1}(k)-1} w'_{v,\pi_v(i)} \tag{8.2}$$

$$\Leftrightarrow y_k = 1 \wedge \sum_{i=1}^{\pi_v^{-1}(k)} w'_{v,\pi_v(i)} = 0. \tag{8.3}$$

We can conclude from $y_k = 1$ and the feasibility of (y, x, z, w) that $\sum_{i=1}^{\pi_v^{-1}(k)} w_{v,\pi_v(i)} > 0$, which necessitates that $\sum_{i=1}^{\pi_v^{-1}(k)} w'_{v,\pi_v(i)} > 0$. But this is a contradiction to assumption (8.3) and completes the proof that (y, x, z, w') satisfies constraints (8.1h). The validity of inequalities (8.1i) can be shown analogously. Hence, (y, x, z, w') is a feasible solution with the same objective value as (y, x, z, w). □

Based on this insight, we can now show that if (Det) is feasible, there always exists an optimal solution in which all unsteerable patient demands target their closest operated treatment facility.

Lemma 8.7. *Let (y, x, z, w) be a feasible solution to (Det) with $\sum_{k \in N(v)} w_{vk} \leq 1$ for every demand origin $v \in V$. Then for all $v \in V$ and $k \in N(v)$, $w_{vk} = 1$ if and only if k is v's closest operating treatment facility, i.e., $k = k_v^{\min}(y)$.*

Proof. Let $v \in V$ be a demand origin with closest operating treatment facility $k_v^{\min}(y) = \pi_v(i) \in N(v)$, $i \in \{1, \ldots, |N(v)|\}$. As all facilities $\pi_v(j)$ for $j < i$ are unoperated MMU sites by the definition of i, i.e., $\pi_v(j) \in N_L(v)$ with $y_{\pi_v(j)} = 0$, we get that $w_{v\pi_v(j)} = 0$ for all $j < i$ by (8.1g). The feasibility of (y, x, z, w) now yields

$$w_{v,k_v^{\min}(y)} \geq 1 - \sum_{j=1}^{i-1} w_{v,\pi_v(j)} = 1.$$

Conversely, let $w_{vk} = 1$ for some $v \in V$ and some $k \in N(v)$. Assume k is not the closest operating treatment facility to v, i.e., for the closest operated treatment facility $k_v^{\min}(y) = \pi_v(i) \in N(v)$ holds $i < \pi_v^{-1}(k)$. As we have $\sum_{j=1}^{N(v)} w_{v,\pi_v(j)} \leq 1$, it directly follows that $w_{v\pi_v(j)} = 0$ for all $j \neq \pi_v^{-1}(k)$. However, this implies that

$$w_{v,k_v^{\min}(y)} = 0 < 1 = 1 - \sum_{j=1}^{i-1} w_{v,\pi_v(j)},$$

which yields a violation of (8.1h) or (8.1i) for v and $k_v^{\min}(y)$, which in turn is a contradiction to the feasibility of (y, x, z, w). □

Concerning the steerable patient demand, we have to show that if (Det) is feasible, there always exists an optimal solution in which no more than the steerable patient demand $d_v \in \mathbb{N}$ originates in each demand origin $v \in V$.

Lemma 8.8. *Given a feasible solution (y, x, z, w) to (Det), we can compute a feasible solution (y, x, z', w) to (Det) with the same objective value and $\sum_{k \in N(v)} z'_{vk} = d_v$ for all $v \in V$ in linear time.*

Proof. If for some $v \in V$ it holds that $d_v^+ \coloneqq \sum_{k \in N(v)} z_{vk} - d_v > 0$, we can arbitrarily reduce the assigned steerable patient demands by d_v^+ without violating constraints (8.1d) or (8.1e), e.g., by setting

$$z'_{vk} \coloneqq z_{vk} - \min \left\{ z_{vk},\, \max \left\{ 0,\, d_v^+ - \sum_{i=1}^{\pi_v^{-1}(k)-1} z_{v\pi_v(i)} \right\} \right\}.$$

□

The correctness of (Det) for the SMMU is now immediate.

Theorem 8.9. (Det) *is an integer linear formulation for the SMMU.*

Proof. Given an optimal solution (y, x, z, w) to (Det), a strategic MMU operation plan m^{S} of minimum cost can be defined via $m^{\mathrm{S}}_\ell \coloneqq x_\ell$ for all $\ell \in L$. By Lemma 8.8, we can assume w.l.o.g. that $\sum_{k \in N(v)} z_{vk} = d_v$ and thus $\{f_v\}_{v \in V}$ defined via $f_v(k) \coloneqq z_{vk}$ for all $k \in N(v)$ induces an assignment of the steerable patient demands. The feasibility of m^{S} now follows directly from Lemmata 8.6 and 8.7 for the assignment $\{f_v\}_{v \in V}$. □

The SMMU determines the set up MMU operation sites, the number of weekly sessions operated per site, as well as the assignment of the steerable patient demands to the treatment facilities. In the subsequent section, we assume that the actual patient demands are uncertain and reveal themselves only after we have fixed our decisions regarding the set up sites and operated MMU sessions. Within this setting, it is no longer expedient to determine one fixed assignment of the steerable patient demands that is feasible for all demand realizations. Instead, we model a flexible assignment of the steerable demands that can be adjusted once the actual demands are known.

Adding assignment variables for every potential demand realization to (Det) leads to a huge model extension that is likely to be computationally intractable. We therefore propose an alternative formulation for the deterministic SMMU that considers the steerable patient demands in a subproblem and is thus much better suited to uncertain patient demands. To that end, we extend the results in Krumke et al. (2019) and employ a Benders decomposition approach to (Det) that decides and fixes the strategic MMU operation plan in the master problem and only checks the plan's feasibility in the subproblem. More precisely, we choose

our first stage variables to be y, x, and w and our second stage variables to be z. The resulting equivalent reformulation of (Det) then reads

$$
\begin{aligned}
(\mathrm{MP}) \quad \min_{y,\,x,\,w} \quad & \sum_{\ell \in L} c_\ell \, y_\ell + \sum_{\ell \in L} \hat{c} \, x_\ell && && (8.4a)\\
\text{s.t.} \quad & (\mathrm{SP})(y, x, w) \text{ is feasible} && && (8.4b)\\
& x_\ell \leq b_\ell \, y_\ell && \forall \ell \in L && (8.4c)\\
& \sum_{k \in N(v)} w_{vk} \geq 1 && \forall v \in V && (8.4d)\\
& w_{v\ell} \leq y_\ell && \forall v \in V,\, \forall \ell \in N_L(v) && (8.4e)\\
& w_{v\ell} \geq y_\ell - \sum_{i=1}^{\pi_v^{-1}(\ell)-1} w_{v,\pi_v(i)} && \forall v \in V,\, \forall \ell \in N_L(v) && (8.4f)\\
& w_{vp} \geq 1 - \sum_{i=1}^{\pi_v^{-1}(p)-1} w_{v,\pi_v(i)} && \forall v \in V,\, \forall p \in N_P(v) && (8.4g)\\
& x_\ell \in \mathbb{N},\ y_\ell \in \{0,1\} && \forall \ell \in L && (8.4h)\\
& w_{vk} \in \{0,1\} && \forall v \in V,\, \forall k \in N(v), && (8.4i)
\end{aligned}
$$

where $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ denotes the Benders subproblem for fixed first-stage decisions $\hat{y}$, $\hat{x}$, and $\hat{w}$, which is defined as

$$
\begin{aligned}
(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w}) \quad \min_{z} \quad & 0 && && (8.5a)\\
\text{s.t.} \quad & \sum_{k \in N(v)} z_{vk} \geq d_v && \forall v \in V && (8.5b)\\
& \sum_{v \in N(\ell)} z_{v\ell} \leq \hat{b} \, \hat{x}_\ell - \sum_{v \in N(\ell)} u_v \, \hat{w}_{v\ell} && \forall \ell \in L && (8.5c)\\
& \sum_{v \in N(p)} z_{vp} \leq \bar{b}_p - \sum_{v \in N(p)} u_v \, \hat{w}_{vp} && \forall p \in P && (8.5d)\\
& z_{vk} \in \mathbb{N} && \forall v \in V,\, \forall k \in N(v). && (8.5e)
\end{aligned}
$$

Next, we investigate the feasibility of the Benders subproblem $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ to derive Benders feasibility cuts which enforce constraint (8.4b). Let us first note, that the constraint matrix of $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ is totally unimodular.

Lemma 8.10. *The constraint matrix of* $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ *is totally unimodular.*

Proof. All entries in the constraint matrix of $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ are in $\{0, 1, -1\}$. Moreover, in every column of the constraint matrix at most one entry in the rows corresponding to constraints (8.5b) takes the value 1 and at most one entry in the rows corresponding to constraints (8.5c) and (8.5d) takes the value -1. Thus, by adding all rows of the constraint matrix into the same partitioning set, the total unimodularity of the constraint matrix of $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ follows directly from the theorem of Hoffman and Gale (Heller and Tompkins, 1956). □

As the right hand sides of the constraints in $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ are integral for all integer feasible first-stage decisions $\hat{y}$, $\hat{x}$, and $\hat{w}$, Lemma 8.10 and Cramer's rule yield that the LP-relaxation of $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ has an integer solution whenever it is feasible. Thus, we can relax constraints (8.5e) and get the following result.

Corollary 8.11. *The Benders subproblem* $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ *is feasible if and only if its LP-relaxation* $(\mathrm{SP}_{\mathrm{LP}})(\hat{y}, \hat{x}, \hat{w})$ *is feasible.*

In order to obtain our Benders feasibility cuts, we exploit the fact that $(\mathrm{SP}_{\mathrm{LP}})(\hat{y}, \hat{x}, \hat{w})$ is the decision version of a maximum flow problem. To ease notation, we define the *residual treatment capacity* of a treatment facility in the Benders subproblem as the treatment capacity that remains after the assignment of the unsteerable patient demands is fixed, i.e.,

$$\gamma_\ell \coloneqq \hat{b}\,\hat{x}_\ell - \sum_{v \in N(\ell)} u_v\,\hat{w}_{v\ell} \qquad \forall \ell \in L,$$
$$\gamma_p \coloneqq \bar{b}_p - \sum_{v \in N(p)} u_v\,\hat{w}_{vp} \qquad \forall p \in P.$$

Throughout this thesis, we will always assume that the residual treatment capacities are non-negative.

Assumption 1. For all feasible solutions $(\hat{y}, \hat{x}, \hat{w})$ to the master problem (MP) without constraint (8.4b), it holds that the residual capacities $\gamma_k \geq 0$ for all $k \in L \cup P$.

Obviously, Assumption 1 does not hold in general. However, we can easily enforce Assumption 1 by adding additional constraints to (MP). As this does not offer new insights but only complicates our formulation, we cover the explicit enforcement of Assumption 1 in Appendix A.1.

The flow network corresponding to $(\mathrm{SP}_{\mathrm{LP}})(\hat{y}, \hat{x}, \hat{w})$ is now constructed as follows. Let G be the directed graph with vertex set $V(G) = \{s\} \cup V \cup (L \cup P) \cup \{t\}$ and arc set $E(G) = E_1 \cup E_2 \cup E_3$, where

$$E_1 \coloneqq \{(s, v) \colon v \in V\},$$
$$E_2 \coloneqq \{(k, t) \colon k \in L \cup P\},$$
$$E_3 \coloneqq \{(v, k) \colon v \in V, k \in N(v)\}.$$

We set the capacities of arcs $e_1 = (s, v) \in E_1$ to $\mu(e_1) \coloneqq d_v$ and the capacities of arcs $e_2 = (k, t) \in E_2$ to $\mu(e_2) \coloneqq \gamma_k$. The capacities of all arcs $e_3 \in E_3$ are set to $\mu(e_3) \coloneqq \infty$. Note, that this choice of arc capacities requires Assumption 1 to hold as we might otherwise end up with negative arc capacities. An example of the constructed network (G, μ, s, t) can be found in Figure 8.2. The following now holds true.

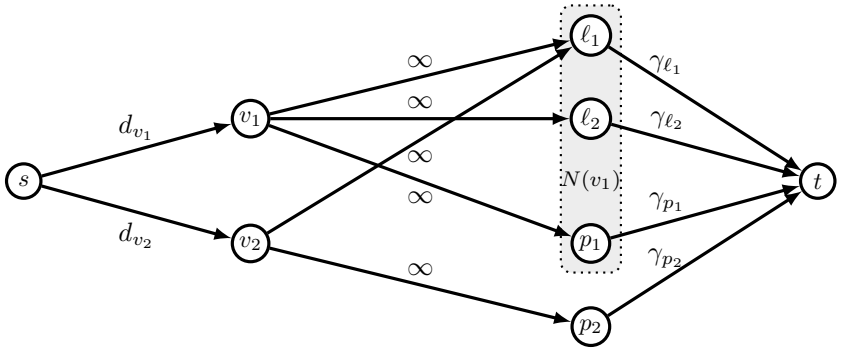


Fig. 8.2.: Example for the network (G, μ, s, t) constructed for Lemma 8.12.

Lemma 8.12. *The Benders subproblem* $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ *is feasible if and only if the maximum* s*-*t*-flow in the network* (G, μ, s, t) *has a flow value of at least* $D := \sum_{v \in V} d_v$.

Proof. Let $f \colon E(G) \to \mathbb{R}_+$ be an s-t flow in (G, μ, s, t) of value value$(f) \geq D$. We define a solution for $(\mathrm{SP}_{\mathrm{LP}})(\hat{y}, \hat{x}, \hat{w})$ by setting $z_{vk} := f((v,k))$ for all $v \in V$, $k \in N(v)$ and show that z is feasible. As the s-t cut induced by $S := \{s\}$ has capacity $\mu(\delta^+(S)) = D$, it follows that value$(f) = D$. It must thus hold for all arcs $(s,v) \in E_1$ that $f((s,v)) = d_v$ and by flow-conservation we get that for all $v \in V$

$$\sum_{k \in N(v)} z_{vk} = \sum_{k \in N(v)} f((v,k)) = \sum_{e \in \delta^+(v)} f(e) = \sum_{e \in \delta^-(v)} f(e) = f((s,v)) = d_v.$$

Moreover, for $k \in L \cup P$ we have that

$$\sum_{v \in N(k)} z_{vk} = \sum_{v \in N(k)} f((v,k)) = \sum_{e \in \delta^-(k)} f(e) = \sum_{e \in \delta^+(k)} f(e) = f((k,t)) \leq \mu((k,t)) = \gamma_k.$$

As a result, z defines a feasible solution for $(\mathrm{SP}_{\mathrm{LP}})(\hat{y}, \hat{x}, \hat{w})$ which implies the feasibility of $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ by Corollary 8.11. The converse direction can be shown analogously. □

We can now combine our intermediate results to derive Benders feasibility cuts by the application of the max-flow min-cut theorem (Ahuja et al., 1993).

Theorem 8.13. *The Benders subproblem* $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ *is feasible if and only if*

$$\sum_{v \in U} d_v + \sum_{k \in N(U)} \sum_{v \in N(k)} u_v \, \hat{w}_{vk} \leq \sum_{\ell \in N_L(U)} \hat{b} \, \hat{x}_\ell + \sum_{p \in N_P(U)} \bar{b}_p \qquad \forall U \subseteq V, \tag{8.6}$$

where $N(U) := \bigcup_{v \in U} N(v)$, $N_L(U) := N(U) \cap L$, *and* $N_P(U) := N(U) \cap P$ *for* $U \subseteq V$.

Proof. First, let us note that (8.6) can be equivalently reformulated as

$$\begin{aligned}
&\sum_{v\in U} d_v + \sum_{k\in N(U)} \sum_{v\in N(k)} u_v\, \hat{w}_{vk} \leq \sum_{\ell\in N_L(U)} \hat{b}\, \hat{x}_\ell + \sum_{p\in N_P(U)} \bar{b}_p && \forall U \subseteq V \\
\Leftrightarrow &\sum_{v\in U} d_v \leq \sum_{\ell\in N_L(U)} \left(\hat{b}\, \hat{x}_\ell - \sum_{v\in N(\ell)} u_v\, \hat{w}_{v\ell}\right) + \sum_{p\in N_P(U)} \left(\bar{b}_p - \sum_{v\in N(p)} u_v\, \hat{w}_{vp}\right) && \forall U \subseteq V \\
\Leftrightarrow &\sum_{v\in U} d_v \leq \sum_{k\in N(U)} \gamma_k && \forall U \subseteq V. \qquad (8.7)
\end{aligned}$$

Moreover, by Lemma 8.12 and the max-flow-min-cut theorem, $(\mathrm{SP})(\hat{y}, \hat{x}, \hat{w})$ is feasible if and only if every s-t cut in the network (G, μ, s, t) induced by $S \subsetneq V(G)$ with $s \in S$, $t \notin S$ has capacity $\mu(\delta^+(S)) \geq D$. Hence, it suffices to show that inequalities (8.7) hold if and only if every s-t cut induced by $S \subsetneq V(G)$ has capacity $\mu(\delta^+(S)) \geq D$.

Assume that the inequalities (8.7) hold. All s-t cuts $\delta^+(S) \subseteq E(G)$ containing arcs from E_3 have infinite capacity and obviously satisfy $\mu(\delta^+(S)) \geq D$. Hence, let $\delta^+(S)$ with $s \in S$, $t \notin S$ be an s-t cut in G of finite capacity and define $U \coloneqq S \cap V$. Then $N(U) \subseteq S$ as otherwise $\delta^+(S) \cap E_3 \neq \emptyset$. Consequently we have that

$$\mu(\delta^+(S)) \geq \sum_{k\in N(U)} \gamma_k + \sum_{v\in V\setminus U} d_v \geq \sum_{v\in U} d_v + \sum_{v\in V\setminus U} d_v = D.$$

Conversely, assume that $\mu(\delta^+(S)) \geq D$ for all $S \subsetneq V(G)$ with $s \in S$, $t \notin S$. Let $U \subseteq V$ and define $S \coloneqq \{s\} \cup U \cup N(U)$. Then obviously $s \in S$ and $t \notin S$ and we get by our assumption that

$$\begin{aligned}
\mu(\delta^+(S)) = \sum_{k\in N(U)} \gamma_k + \sum_{v\in V\setminus U} d_v \geq \sum_{v\in V} d_v \\
\Leftrightarrow \sum_{k\in N(U)} \gamma_k \geq \sum_{v\in U} d_v.
\end{aligned}$$

□

As a result of Theorem 8.13, we can obtain a linear formulation of our Benders master problem (MP) by substituting constraint (8.4b) with the Benders feasibility cuts (8.6). We refer to the resulting formulation of the SMMU as $(\mathrm{Det\text{-}B})$.

Corollary 8.14. $(\mathrm{Det\text{-}B})$ *is an integer linear formulation for the SMMU.*

For all integer first-stage solutions, the Benders feasibility cuts (8.6) can be separated in polynomial time by computing a minimum s-t cut in the network (G, μ, s, t) as described above. Alternatively, one can separate the cuts by solving the dual of the Benders subproblem $(\mathrm{SP_{LP}})(\hat{y}, \hat{x}, \hat{w})$ that we describe in Appendix A.2. Appendix A.2 furthermore shows that the separation problem for $(\mathrm{Det\text{-}B})$ is trivial if we only consider unsteerable patient demands due to Assumption 1.

The next section considers the SMMU with uncertain patient demands. As our main interest lies on the setup of operation sites and operation of MMU sessions, it suffices to guarantee the existence of a feasible assignment of the steerable patient demands. Thus, we restrict ourselves to the Benders formulation $(\mathrm{Det\text{-}B})$ of the SMMU in the following and note that such an assignment can be determined by a single maximum flow computation as a result of Lemma 8.12.

8.2 Integration of Demand Uncertainties

Up to this point, we considered the SMMU in a deterministic setting. That is we assume that all input data is precisely known, in particular, we assume that the weekly steerable and unsteerable patient demand at each demand origin $v \in V$ can be described by deterministic nominal values $d_v \in \mathbb{N}$ and $u_v \in \mathbb{N}$, respectively. Clearly, this assumption does not hold in reality as a patient's need to see a primary care physician is subject to fluctuation. As a result, strategic MMU operation plans that are feasible with respect to the nominal patient demands may be infeasible in real-life operation (Ben-Tal and Nemirovski, 2000). To address this issue, we model the weekly patient demands at each demand origin as random variables. Specifically, we assume that the steerable patient demand at each demand origin $v \in V$ can be described by an independent random variable ξ_v that takes values in $\{\alpha_v, \alpha_v + 1, \ldots, \beta_v\}$, where $\alpha_v, \beta_v \in \mathbb{N}$ with $\alpha_v \leq \beta_v$ are the respective steerable lower and upper bounds. Analogously, we assume that the unsteerable patient demand at each demand origin $v \in V$ can be described by an independent random variable η_v that takes values in $\{\sigma_v, \sigma_v + 1, \ldots, \tau_v\}$, where $\sigma_v, \tau_v \in \mathbb{N}$ with $\sigma_v \leq \tau_v$ are the respective unsteerable lower and upper bounds.

To extend the SMMU to uncertain patient demands, we employ the concept of robust optimization (Ben-Tal et al., 2009; Gabrel et al., 2014). The core principle of robust optimization is to strive for solutions that are, to some extent, immune to variations in the input data. This is achieved by hedging solutions against a subset of all possible realizations of the uncertain parameters which are represented by so-called *uncertainty sets*.

Under our model of data uncertainty, the set of all possible realizations of the steerable and unsteerable patient demands are given by $\Xi \coloneqq \{\xi \in \mathbb{N}^V : \alpha_v \leq \xi_v \leq \beta_v \ \forall v \in V\}$ and $H \coloneqq \{\eta \in \mathbb{N}^V : \sigma_v \leq \eta_v \leq \tau_v \ \forall v \in V\}$, respectively. The *robust strategic planning problem for MMUs* then asks for a strategic MMU operation plan of minimum cost that is feasible for every pair of patient demand realizations $(\xi, \eta) \in \mathcal{U}_1 \times \mathcal{U}_2$, where $\mathcal{U}_1 \subseteq \Xi$ is an uncertainty set of the steerable patient demands and $\mathcal{U}_2 \subseteq H$ is an uncertainty set of the unsteerable patient demands. To formalize this, we extend the notion of a feasible strategic MMU operation plan to the robust setting with uncertain patient demands.

Definition 8.15. A strategic MMU operation plan m^{S} is *robust feasible* if m^{S} is feasible for the deterministic SMMU with nominal patient demands $d = \xi$ and $u = \eta$ for every pair of patient demand realizations $(\xi, \eta) \in \mathcal{U}_1 \times \mathcal{U}_2$.

Note, that the number of sessions operated at each site is fixed in a robust feasible MMU operation plan which induces a fixed assignment of the unsteerable demands that is independent of the realization $\eta \in \mathcal{U}_2$. However, the assignment of the steerable patient demands is variable and can be adapted for any realization $\xi \in \mathcal{U}_1$. We can now use the notion of a robust feasible strategic MMU operation plan to provide a formal definition of the robust strategic planning problem for MMUs.

Definition 8.16 (rSMMU)**.** Let the potential MMU operation sites $\ell \in L$ with setup costs $c_\ell \in \mathbb{N}$ and weekly session capacities $b_\ell \in \mathbb{N}$ be given. Moreover, let $p \in P$ be the existing practices with weekly treatment capacities $\bar{b}_p \in \mathbb{N}$ and $v \in V$ be the patient demand origins with consideration sets $N(v) \subseteq L \cup P$. The uncertain weekly steerable and unsteerable demands are described by the uncertainty sets $\mathcal{U}_1 \subseteq \Xi$ and $\mathcal{U}_2 \subseteq H$, respectively. Then, the *robust strategic planning problem for MMUs* (rSMMU) asks for a robust feasible strategic MMU operation plan of minimum cost, where every operated MMU session induces the cost $\hat{c} \in \mathbb{N}$ and yields a weekly treatment capacity $\hat{b} \in \mathbb{N}$.

Obviously, the rSMMU is a generalization of the SMMU and thus the problem's strong $\mathcal{NP}$-hardness follows immediately from Theorem 8.5.

Corollary 8.17. *The rSMMU is strongly $\mathcal{NP}$-hard.*

To obtain a formulation for the rSMMU, we consider the robust counterpart of the formulation (Det-B) for the deterministic SMMU defined as

$$
\begin{aligned}
(\text{Rob-B}) \quad \min_{y,\,x,\,w} \quad & \sum_{\ell \in L} c_\ell\, y_\ell + \sum_{\ell \in L} \hat{c}\, x_\ell && && \text{(8.8a)}\\
\text{s.t.} \quad & \max_{\xi \in \mathcal{U}_1} \sum_{v \in U} \xi_v + \max_{\eta \in \mathcal{U}_2} \sum_{k \in N(U)} \sum_{v \in N(k)} \eta_v\, w_{vk} && && \\
& \qquad \leq \sum_{\ell \in N_L(U)} \hat{b}\, x_\ell + \sum_{p \in N_P(U)} \bar{b}_p && \forall U \subseteq V && \text{(8.8b)}\\
& x_\ell \leq b_\ell\, y_\ell && \forall \ell \in L && \text{(8.8c)}\\
& \sum_{k \in N(v)} w_{vk} \geq 1 && \forall v \in V && \text{(8.8d)}\\
& w_{v\ell} \leq y_\ell && \forall v \in V,\, \forall \ell \in N_L(v) && \text{(8.8e)}\\
& w_{v\ell} \geq y_\ell - \sum_{i=1}^{\pi_v^{-1}(\ell)-1} w_{v,\pi_v(i)} && \forall v \in V,\, \forall \ell \in N_L(v) && \text{(8.8f)}\\
& w_{vp} \geq 1 - \sum_{i=1}^{\pi_v^{-1}(p)-1} w_{v,\pi_v(i)} && \forall v \in V,\, \forall p \in N_P(v) && \text{(8.8g)}\\
& x_\ell \in \mathbb{N},\ y_\ell \in \{0,1\} && \forall \ell \in L && \text{(8.8h)}\\
& w_{vk} \in \{0,1\} && \forall v \in V,\, \forall k \in N(v). && \text{(8.8i)}
\end{aligned}
$$

In this formulation, inequalities (8.8b) correspond to the robust Benders feasibility cuts, constraints (8.8c) enforce the session capacity at each setup site, and inequalities (8.8d)–(8.8g) ensure that unsteerable patient demands target their closest considered operating treatment facility. We show that (Rob-B) is a formulation for the rSMMU.

Theorem 8.18. (Rob-B) *is an integer formulation for the rSMMU.*

Proof. Given an optimal solution (y, x, w) to (Rob-B), a strategic MMU operation plan m^{S} of minimum cost can be defined via $m^{\mathrm{S}}_\ell := x_\ell$ for all $\ell \in L$. As (y, x, w) satisfies constraints (8.8b), it follows that for every pair of patient demand realizations $(\hat{\xi}, \hat{\eta}) \in \mathcal{U}_1 \times \mathcal{U}_2$ and every $U \subseteq V$ we have that

$$\begin{aligned}\sum_{v \in U} \hat{\xi}_v + \sum_{k \in N(U)} \sum_{v \in N(k)} \hat{\eta}_v \, w_{vk} &\leq \max_{\xi \in \mathcal{U}_1} \sum_{v \in U} \xi_v + \max_{\eta \in \mathcal{U}_2} \sum_{k \in N(U)} \sum_{v \in N(k)} \eta_v \, w_{vk} \\ &\leq \sum_{\ell \in N_L(U)} \hat{b} \, x_\ell + \sum_{p \in N_P(U)} \bar{b}_p.\end{aligned}$$

Thus, the robust feasibility of m^{S} follows directly from Theorem 8.13. □

Formulation (Rob-B) is in general non-linear due to constraints (8.8b). However, for certain choices of the uncertainty sets $\mathcal{U}_1 \subseteq \Xi$ and $\mathcal{U}_2 \subseteq H$ we can show that (8.8b) can be reformulated in a linear way. There are various concepts of defining uncertainty sets; see, e.g., Bertsimas and Sim (2003), Bertsimas and Sim (2004), Kouvelis and Yu (1996), Kasperski (2008), and Kasperski and Zieliński (2016). The first of these setting we consider, is the complete protection against uncertainties in the patient demands, i.e., the rSMMU with uncertainty sets $\mathcal{U}_1 = \Xi$ and $\mathcal{U}_2 = H$. This setting is known as *interval uncertainty* (Ben-Tal et al., 2009; Soyster, 1973) and allows us to reformulate (8.8b) as

$$\sum_{v \in U} \beta_v + \sum_{k \in N(U)} \sum_{v \in N(k)} \tau_v \, w_{vk} \leq \sum_{\ell \in N_L(U)} \hat{b} \, x_\ell + \sum_{p \in N_P(U)} \bar{b}_p \quad \forall U \subseteq V. \tag{8.8b′}$$

That is, we can reduce the rSMMU for this particular choice of uncertainty sets to the deterministic SMMU with worst-case nominal patient demands $d_v = \beta_v$ and $u_v = \tau_v$ for all $v \in V$. We refer to the resulting formulation of the rSMMU with interval uncertainty sets as (RobI-B). This approach is known as the method of Soyster (Soyster, 1973) and generally entails prohibitive operation cost as a result of the method's conservatism.

To alleviate this drawback, Bertsimas and Sim (2004) introduced *budgeted uncertainty sets* that restrict the deviations in the uncertain input data through a budget parameter. The choice of this budget parameter allows for a trade-off between robustness and operation cost of the obtained solutions. In the following, we consider an adaptation of budgeted uncertainty sets that contains all patient demand realizations in which the total (un-)steerable

patient demand is bounded by the parameter $\Gamma_1 \in \mathbb{N}$ ($\Gamma_2 \in \mathbb{N}$). For the steerable patient demands, these realizations can be represented by the uncertainty set

$$\mathcal{U}_1^\Gamma \coloneqq \left\{ \xi \in \mathbb{N}^V : \alpha_v \leq \xi_v \leq \beta_v \; \forall v \in V, \; \sum_{v \in V} \xi_v \leq \Gamma_1 \right\}.$$

For the unsteerable patient demands, we analogously obtain the uncertainty set

$$\mathcal{U}_2^\Gamma \coloneqq \left\{ \eta \in \mathbb{N}^V : \sigma_v \leq \eta_v \leq \tau_v \; \forall v \in V, \; \sum_{v \in V} \eta_v \leq \Gamma_2 \right\}.$$

To ensure that the uncertainty sets $\mathcal{U}_1^\Gamma$ and $\mathcal{U}_2^\Gamma$ are non-empty, we require that $\sum_{v \in V} \alpha_v \leq \Gamma_1$ and $\sum_{v \in V} \sigma_v \leq \Gamma_2$. Moreover, we can assume w.l.o.g. that $\Gamma_1 \leq \sum_{v \in V} \beta_v$ and $\Gamma_2 \leq \sum_{v \in V} \tau_v$ as we otherwise always have $\mathcal{U}_1^\Gamma = \Xi$ and $\mathcal{U}_2^\Gamma = H$.

For the remainder of this section, we consider the rSMMU with the budgeted uncertainty sets $\mathcal{U}_1^\Gamma$ and $\mathcal{U}_2^\Gamma$ and devise an integer linear formulation, which is subsequently solved by constraint generation. To that end, we show that (8.8b) can be linearized for this particular choice of uncertainty sets.

Considering the non-linear part in (8.8b) corresponding to the steerable patient demands, the linear reformulation is straightforward as

$$\max_{\xi \in \mathcal{U}_1^\Gamma} \sum_{v \in U} \xi_v = \min \left\{ \sum_{v \in U} \beta_v, \; \Gamma_1 - \sum_{v \in V \setminus U} \alpha_v \right\} \tag{8.9}$$

which is simply a constant for fixed $U \subseteq V$.

For the non-linear part in (8.8b) corresponding to the unsteerable patient demands, we can obtain a linear reformulation through LP duality. By the definition of $\mathcal{U}_2^\Gamma$, we can formulate the inner maximization problem

$$\max_{\eta \in \mathcal{U}_2^\Gamma} \sum_{k \in N(U)} \sum_{v \in N(k)} \eta_v \, \hat{w}_{vk}$$

for fixed $U \subseteq V$ and fixed assignment of the unsteerable demands $\hat{w}_{vk} \in \{0, 1\}$ for all $v \in V$ and $k \in N(v)$ via the following integer linear program:

$$(\mathrm{P}^U)(\hat{w}) \quad \max_{\eta} \quad \sum_{k \in N(U)} \sum_{v \in N(k)} \eta_v \, \hat{w}_{vk} \tag{8.10a}$$

$$\text{s.t.} \quad \eta_v \leq \tau_v \qquad \forall v \in V \tag{8.10b}$$

$$-\eta_v \leq -\sigma_v \qquad \forall v \in V \tag{8.10c}$$

$$\sum_{v \in V} \eta_v \leq \Gamma_2 \tag{8.10d}$$

$$\eta_v \in \mathbb{N} \qquad \forall v \in V. \tag{8.10e}$$

The problem $(\mathrm{P}^U)(\hat{w})$ is feasible and bounded as we assumed $\Gamma_2 \geq \sum_{v \in V} \sigma_v$. Moreover, we can show that the constraint matrix of $(\mathrm{P}^U)(\hat{w})$ is totally unimodular.

Lemma 8.19. *The constraint matrix of* $(\mathrm{P}^U)(\hat{w})$ *is totally unimodular.*

Proof. The unit rows of the constraint matrix corresponding to constraints (8.10b) and (8.10c) are irrelevant to the total unimodularity and do not have to be considered (Nemhauser and Wolsey, 2014a). Thus, we end up with a vector of ones corresponding to constraint (8.10d) which is obviously totally unimodular as each square submatrix has determinant one. □

By our choice of parameters, the right hand sides of the constraints in $(\mathrm{P}^U)(\hat{w})$ are integral. Thus, the polyhedron of the LP-relaxation $(\mathrm{P}^U_{\mathrm{LP}})(\hat{w})$ is integral and we can relax the integrality constraint (8.10e) as the optimal solution values of $(\mathrm{P}^U_{\mathrm{LP}})(\hat{w})$ and $(\mathrm{P}^U)(\hat{w})$ coincide. The dual problem of $(\mathrm{P}^U_{\mathrm{LP}})(\hat{w})$ is given by

$$
\begin{aligned}
(\mathrm{D}^U_{\mathrm{LP}})(\hat{w}) \quad & \min_{\varepsilon,\, \kappa,\, \rho} \quad \sum_{v \in V} (\tau_v \varepsilon_v - \sigma_v \kappa_v) + \Gamma_2 \rho \\
& \text{s.t.} \quad \varepsilon_v - \kappa_v + \rho \geq \sum_{k \in N(U) \cap N(v)} \hat{w}_{vk} \quad \forall v \in V \\
& \qquad\quad \varepsilon_v, \kappa_v, \rho \geq 0 \qquad\qquad\qquad\qquad\quad \forall v \in V.
\end{aligned}
$$

Strong duality states that the optimal solution values of $(\mathrm{P}^U_{\mathrm{LP}})(\hat{w})$ and $(\mathrm{D}^U_{\mathrm{LP}})(\hat{w})$ coincide. Hence, every feasible solution of $(\mathrm{D}^U_{\mathrm{LP}})(\hat{w})$ yields an upper bound on the optimal solution value of $(\mathrm{P}^U)(\hat{w})$. Combined with the observations in Bertsimas and Sim (2004), we can now reformulate (8.8b) for the budgeted uncertainty sets $\mathcal{U}_1^\Gamma$ and $\mathcal{U}_2^\Gamma$ via the following set of constraints:

$$
\min \left\{ \sum_{v \in U} \beta_v,\ \Gamma_1 - \sum_{v \in V \setminus U} \alpha_v \right\} + \sum_{v \in V} \left(\tau_v \varepsilon_v^U - \sigma_v \kappa_v^U \right) + \Gamma_2 \rho^U
$$

$$
\leq \sum_{\ell \in N_L(U)} \hat{b} x_\ell + \sum_{p \in N_P(U)} \bar{b}_p \qquad \forall U \subseteq V \tag{8.11}
$$

$$
\varepsilon_v^U - \kappa_v^U + \rho^U \geq \sum_{k \in N(U) \cap N(v)} w_{vk} \qquad \forall v \in V, \forall U \subseteq V \tag{8.12}
$$

$$
\varepsilon_v^U, \kappa_v^U, \rho^U \geq 0 \qquad \forall v \in V, \forall U \subseteq V. \tag{8.13}
$$

We refer to the resulting formulation of the rSMMU with budgeted uncertainty sets as $(\mathrm{Rob}\Gamma\text{-B})$. Formulation $(\mathrm{Rob}\Gamma\text{-B})$ is an integer linear program with an exponential number of constraints. To solve it, we apply constraint generation, i.e., we consider $(\mathrm{Rob}\Gamma\text{-B})$ with a subset of the constraints of type (8.11)–(8.13). In particular, we decide on some $\mathscr{U} \subseteq 2^V$

and consider the constraints of type (8.11)–(8.13) only for the subsets of patient demand origins $U \in \mathscr{U}$. This yields a relaxation of $(\mathrm{Rob}\Gamma\text{-B})$ called the *restricted master problem*.

Once an optimal solution $(\hat{y}, \hat{x}, \hat{w}, \hat{\varepsilon}, \hat{\kappa}, \hat{\rho})$ to the restricted master problem induced by $\mathscr{U}$ is known, we need to decide whether $(\hat{y}, \hat{x}, \hat{w}, \hat{\varepsilon}, \hat{\kappa}, \hat{\rho})$ is feasible for the original formulation $(\mathrm{Rob}\Gamma\text{-B})$. To that end, we examine whether there exists a subset $U \subseteq V$ for which the system (8.11)–(8.13) is infeasible. This problem is known as the *separation problem* and can be formalized as follows: Is there a subset $U \subseteq V$ such that the system

$$\begin{aligned}
\min\left\{\sum_{v\in U}\beta_v,\ \Gamma_1 - \sum_{v\in V\setminus U}\alpha_v\right\} + \sum_{v\in V}\left(\tau_v\varepsilon_v^U - \sigma_v\kappa_v^U\right) + \Gamma_2\rho^U & \\
\leq \sum_{\ell\in N_L(U)}\hat{b}\hat{x}_\ell + \sum_{p\in N_P(U)}\bar{b}_p & \\
\varepsilon_v^U - \kappa_v^U + \rho^U \geq \sum_{k\in N(U)\cap N(v)}\hat{w}_{vk} & \qquad \forall v \in V \\
\varepsilon_v^U, \kappa_v^U, \rho^U \geq 0 & \qquad \forall v \in V
\end{aligned}$$

has no solution $(\varepsilon^U, \kappa^U, \rho^U)$, i.e., is infeasible?

By duality and Farkas' lemma (Nemhauser and Wolsey, 2014b), we can equivalently reformulate the separation problem in terms of the original constraints (8.8b): Is there a subset $U \subseteq V$ such that

$$\max_{\xi\in\mathcal{U}_1^\Gamma}\sum_{v\in U}\xi_v + \max_{\eta\in\mathcal{U}_2^\Gamma}\sum_{k\in N(U)}\sum_{v\in N(k)}\eta_v\,\hat{w}_{vk} > \sum_{\ell\in N_L(U)}\hat{b}\,\hat{x}_\ell + \sum_{p\in N_P(U)}\bar{b}_p\ ? \tag{8.14}$$

In the following, we simplify formulation (8.14) of the separation problem even further. To that end, let us recall that for fixed set $U \subseteq V$ we have concluded in (8.9) that for the steerable patient demands holds

$$\max_{\xi\in\mathcal{U}_1^\Gamma}\sum_{v\in U}\xi_v = \min\left\{\sum_{v\in U}\beta_v,\ \Gamma_1 - \sum_{v\in V\setminus U}\alpha_v\right\}. \tag{8.9}$$

Moreover, as the assignment of the unsteerable demands in the separation problem is fixed, we can obtain an analogous result for the unsteerable patient demands. To that end, let $k_v^{\min}(\hat{w}) \in N(v)$ denote the unique treatment facility that is targeted by all unsteerable patient demand originating in $v \in V$, i.e., $\hat{w}_{vk} = 1$ if and only if $k = k_v^{\min}(\hat{w})$. Moreover, let

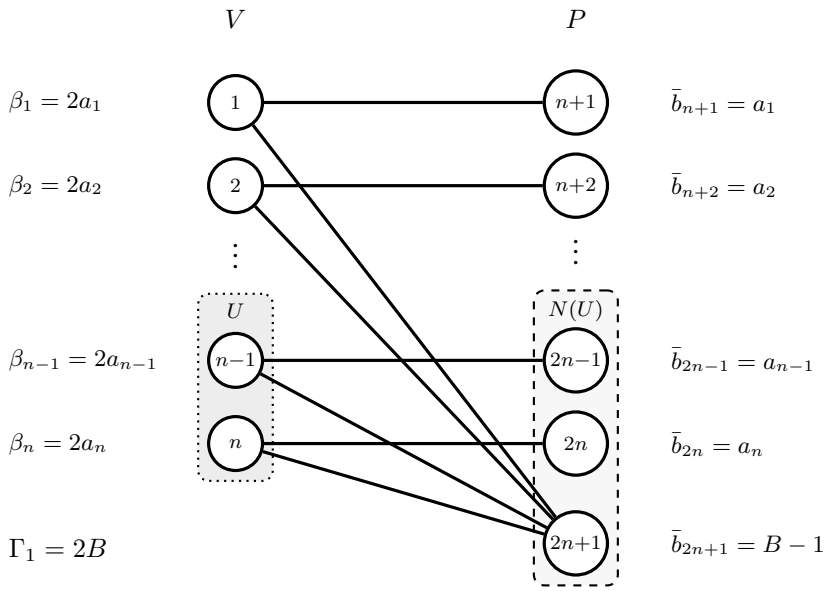


Fig. 8.3.: Constructed separation instance $\mathcal{I}'$ for given subset sum instance $\mathcal{I} = (A, B)$ where consideration sets are encoded by edges in the bipartite graph.

$V^U(\hat{w}) \coloneqq \{v \in V \colon k_v^{\min}(\hat{w}) \in N(U)\}$ denote all demand origins whose unsteerable patient demands target a treatment facility in $N(U) \subseteq L \cup P$. Now we get the following:

$$\begin{aligned} \max_{\eta \in \mathcal{U}_2^\Gamma} \sum_{k \in N(U)} \sum_{v \in N(k)} \eta_v \, \hat{w}_{vk} &= \max_{\eta \in \mathcal{U}_2^\Gamma} \sum_{v \in V^U(\hat{w})} \eta_v \\ &= \min \left\{ \sum_{v \in V^U(\hat{w})} \tau_v,\ \Gamma_2 - \sum_{v \in V \setminus V^U(\hat{w})} \sigma_v \right\}. \end{aligned} \tag{8.15}$$

Substituting (8.9) and (8.15) into (8.14), we obtain the following reformulation of the separation problem: Is there a subset $U \subseteq V$ such that

$$\begin{aligned} \min \left\{ \sum_{v \in U} \beta_v,\ \Gamma_1 - \sum_{v \in V \setminus U} \alpha_v \right\} + \min \left\{ \sum_{v \in V^U(\hat{w})} \tau_v,\ \Gamma_2 - \sum_{v \in V \setminus V^U(\hat{w})} \sigma_v \right\} \\ > \sum_{\ell \in N_L(U)} \hat{b}\, \hat{x}_\ell + \sum_{p \in N_P(U)} \bar{b}_p \, ? \end{aligned} \tag{8.14'}$$

We show, that deciding the separation problem is $\mathcal{NP}$-complete by a reduction from subset sum inspired by the one in Krumke et al. (2019).

Theorem 8.20. *The separation problem for* $(\mathrm{Rob}\Gamma\text{-B})$ *is* $\mathcal{NP}$*-complete.*

Proof. To show the $\mathcal{NP}$-completeness of the separation problem, we perform a reduction from the subset sum problem which is known to be $\mathcal{NP}$-complete (Garey and Johnson, 1979). Let us recall the subset sum problem: Given a finite set $A = \{a_1, \ldots, a_n\} \subseteq \mathbb{N}$ and an integer $B \in \mathbb{N}$, the subset sum problem asks whether there exists a subset $A' \subseteq A$ with $\sum_{a \in A'} a = B$.

Given an instance $\mathcal{I} = (A, B)$ of the subset sum problem, we construct an instance $\mathcal{I}'$ of the separation problem for (RobΓ-B) as follows: Let $V = \{1, \dots, n\}$, $L = \emptyset$, and $P = \{n+1, \dots, 2n, 2n+1\}$. We set $\alpha_v = \sigma_v = \tau_v = 0$ for all $v \in V$, that is we do not consider unsteerable patient demands. Moreover, we set $\beta_v = 2a_v$ for all $v \in V$. Concerning the practices' treatment capacities, we set $\bar{b}_p = a_{p-n}$ for all $p \in P \setminus \{2n+1\}$ and $\bar{b}_{2n+1} = B - 1$. The consideration sets are defined as $N(v) = \{v + n,\ 2n+1\}$ for all $v \in V$ and we choose $\Gamma_1 = 2B$. The construction of $\mathcal{I}'$ is visualized in Figure 8.3.

For our choice of parameters, the separation problem for (RobΓ-B) reduces to: Is there a subset $U \subseteq V$ such that $\min\{\sum_{v\in U} \beta_v,\ \Gamma_1\} > \sum_{p\in N(U)} \bar{b}_p$?

We show that the constructed instance $\mathcal{I}'$ of the separation problem is a Yes-instance if and only if the subset sum instance $\mathcal{I}$ is a Yes-instance.

First, assume that $\mathcal{I}$ is a Yes-instance and let $A' \subseteq A$ with $\sum_{a\in A'} a = B$. Then for $U = \{v \in V : a_v \in A'\}$ it holds that

$$\min\left\{\sum_{v\in U} \beta_v,\ \Gamma_1\right\} = \min\left\{\sum_{a\in A'} 2a,\ 2B\right\} = 2B > 2B - 1 = \sum_{a\in A'} a + B - 1 = \sum_{p\in N(U)} \bar{b}_p$$

which shows that $\mathcal{I}'$ is a Yes-instance.

Conversely, assume that $\mathcal{I}'$ a Yes-instance and let $U \subseteq V$ be a subset of demand origins with $\min\{\sum_{v\in U} \beta_v,\ \Gamma_1\} > \sum_{p\in N(U)} \bar{b}_p$. We show that $A' = \{a_v \in A : v \in U\}$ satisfies $\sum_{a\in A'} a = B$. To that end, we begin by showing that

$$\sum_{v\in U} \beta_v \le \Gamma_1 \Leftrightarrow \sum_{a\in A'} 2a \le 2B \Leftrightarrow \sum_{a\in A'} a \le B. \tag{8.16}$$

Assume the contrary, i.e., that $\sum_{a\in A'} a > B$. Then by our choice of U, we have that

$$\min\left\{\sum_{v\in U} \beta_v,\ \Gamma_1\right\} = \Gamma_1 > \sum_{p\in N(U)} \bar{b}_p \Leftrightarrow 2B > \sum_{a\in A'} a + B - 1 \Leftrightarrow \sum_{a\in A'} a \le B$$

which is a contradiction and thus proves (8.16). By our choice of U, we moreover get

$$\begin{aligned}\min\left\{\sum_{v\in U} \beta_v,\ \Gamma_1\right\} = \sum_{a\in A'} 2a > \sum_{p\in N(U)} \bar{b}_p &\Leftrightarrow \sum_{a\in A'} 2a > \sum_{a\in A'} a + B - 1 \\ &\Leftrightarrow \sum_{a\in A'} a \ge B \end{aligned} \tag{8.17}$$

Combining (8.16) and (8.17), it follows that $\sum_{a\in A'} a = B$ and thus $\mathcal{I}$ is a Yes-instance.

Finally, we remark that the separation problem for (RobΓ-B) is contained in $\mathcal{NP}$ as we can compute all terms in (8.14') for given $U \subseteq V$ in polynomial time. □

Just as in the deterministic setting, the separation problem for $(\mathrm{Rob}\Gamma\text{-B})$ is trivial if we only consider unsteerable demands due to Assumption 1.

To decide the separation problem, we propose an integer linear program based on formulation (8.14'). This formulation requires variables to encode our choice of $U \subseteq V$ as well as the derived sets $N(U) \subseteq L \cup P$ and $V^U(\hat{w}) \subseteq V$. Therefore, we introduce variables $o_v \in \{0,1\}$ that take the value one if demand origin $v \in V$ is in the set U and zero otherwise. Variables $n_k \in \{0,1\}$ take the value one if treatment facility $k \in L \cup P$ is in the consideration set $N(U)$ and zero otherwise. Finally, we introduce variables $r_v \in \{0,1\}$ that take the value one if $v \in V^U(\hat{w})$ and zero otherwise. To linearize the inner minimization problems in (8.14'), we furthermore introduce continuous variables $d_1 \geq 0$ and $d_2 \geq 0$ which attain the value of the respective worst case patient demand for the chosen subset $U \subseteq V$ in an optimal solution. We can now formulate the separation problem as follows:

$$
\begin{aligned}
(\mathrm{Sep}) \quad \max_{d_1, d_2, o, n, r} \quad & d_1 + d_2 - \sum_{\ell \in L} \hat{b}\hat{x}_\ell n_\ell - \sum_{p \in P} \bar{b}_p n_p && && \text{(8.18a)}\\
\text{s.t.} \quad & n_k \geq o_v && \forall v \in V,\ k \in N(v) && \text{(8.18b)}\\
& r_v \leq \sum_{v' \in N(k_v^{\min}(\hat{w}))} o_{v'} && \forall v \in V && \text{(8.18c)}\\
& d_1 \leq \sum_{v \in V} \beta_v o_v && && \text{(8.18d)}\\
& d_1 \leq \Gamma_1 - \sum_{v \in V} \alpha_v (1 - o_v) && && \text{(8.18e)}\\
& d_2 \leq \sum_{v \in V} \tau_v r_v && && \text{(8.18f)}\\
& d_2 \leq \Gamma_2 - \sum_{v \in V} \sigma_v (1 - r_v) && && \text{(8.18g)}\\
& o_v, r_v \in \{0,1\} && \forall v \in V && \text{(8.18h)}\\
& n_k \in \{0,1\} && \forall k \in L \cup P && \text{(8.18i)}\\
& d_1, d_2 \geq 0. && && \text{(8.18j)}
\end{aligned}
$$

Thereby, inequalities (8.18b) enforce that n_k for $k \in L \cup P$ encode the consideration set $N(U)$ and constraints (8.18c) ensure that r_v for $v \in V$ encode $V^U(\hat{w})$. The remaining inequalities (8.18d)–(8.18g) model the reformulated inner minimization problems for the steerable and unsteerable patient demands derived in (8.9) and (8.15), respectively.

Given an optimal solution $(\hat{d}_1, \hat{d}_2, \hat{o}, \hat{n}, \hat{r})$ to (Sep), we can decide the separation problem as follows. If the solution value of $(\hat{d}_1, \hat{d}_2, \hat{o}, \hat{n}, \hat{r})$ is non-positive, it follows that the optimal solution $(\hat{y}, \hat{x}, \hat{w}, \hat{\varepsilon}, \hat{\kappa}, \hat{\rho})$ to the restricted master problem is also an optimal solution to $(\mathrm{Rob}\Gamma\text{-B})$. Otherwise, we get the violating subset $\hat{U} \coloneqq \{v \in V : \hat{o}_v = 1\}$ which is added to $\mathscr{U}$ and we iterate by resolving the restricted master problem.

9 Phase 2: Tactical Planning for MMUs

The previous chapter considered the strategic planning for MMUs in a session-aggregated form which yields a strategic MMU operation plan $m^{\mathrm{S}}\colon L \to \mathbb{N}$ that determines which MMU operation sites are set up and how often these are serviced in the course of one week. Strategic MMU operation plans are not intended for a direct real-world implementation, as they do not contain information regarding what sites are serviced in which sessions of the week. We address this missing link in Phase 2 of P3MMU at the tactical level. To plan the MMU operations in each session of the week, we require some additional formalisms. Let Λ denote the sessions of the week which generally comprise a morning and an afternoon session for every working day of the week, i.e., $\Lambda = \{\mathrm{MON}_{\mathrm{AM}}, \ldots, \mathrm{SAT}_{\mathrm{PM}}\}$. Moreover, we consider the session-expanded potential MMU operation sites $\boldsymbol{L} \coloneqq L \times \Lambda$ where each expanded site $\boldsymbol{\ell} = (\ell, \lambda) \in \boldsymbol{L}$ can be serviced at most once per week. Consequently, we do not allow for parallel MMU operations although we will see later-on that all our results generalize to parallel operations. Finally, we consider the session-expanded practices $\boldsymbol{P} \coloneqq P \times \Lambda$. Using this notation, we can now formally define a tactical MMU operation plan. Recall, that setting up site $\ell \in L$ induces cost $c_\ell \in \mathbb{N}$ while each operated MMU session costs $\hat{c} \in \mathbb{N}$.

Definition 9.1. A *tactical MMU operation plan* is a function $m^{\mathrm{T}}\colon \boldsymbol{L} \to \{0, 1\}$. The *cost* of a tactical MMU operation plan m^{T} is defined by the costs of setting up sites and operating MMU sessions, i.e., we have $c(m^{\mathrm{T}}) \coloneqq \sum_{\ell \in L : \exists \lambda \in \Lambda : m^{\mathrm{T}}_{(\ell,\lambda)} > 0} c_\ell + \sum_{\boldsymbol{\ell} \in \boldsymbol{L}} \hat{c}\, m^{\mathrm{T}}_{\boldsymbol{\ell}}$.

Tactical MMU operation plans define for each session of the week which MMU sites are serviced and thus available to patients. An illustration of a tactical MMU operation plan as it could be used to inform patients can be found in Table 9.1.

We call the serviced sites in session $\lambda \in \Lambda$ for a given tactical MMU operation plan m^{T} the *session service* which we denote by $L_\lambda(m^{\mathrm{T}}) \coloneqq \{\ell \in L : m^{\mathrm{T}}_{(\ell,\lambda)} = 1\}$. As the tactical MMU operation plan to which a session service corresponds is usually clear from context, we abbreviation the session service by $L_\lambda \subseteq L$. The ordered multiset of all session services $\{L_\lambda\}_{\lambda \in \Lambda}$ uniquely defines the tactical MMU operation plan m^{T} and we therefore associate the two with each other to ease notation.

By defining session services, tactical MMU operation plans implicitly determine the minimum number of MMUs $\nu(m^{\mathrm{T}}) \in \mathbb{N}$ required to operate m^{T}, i.e., $\nu(m^{\mathrm{T}}) = \max_{\lambda \in \Lambda} |L_\lambda|$. For example, the tactical MMU operation plan in Table 9.1 would require at least two vehicles ($\nu(m^{\mathrm{T}}) = 2$)

Tab. 9.1.: Illustration of a tactical MMU operation plan.

	Mon		Tue		Wed		Thu		Fri		Sat	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
ℓ_1: train station	✓						✓		✓			
ℓ_2: town hall		✓			✓					✓		
ℓ_3: market square			✓					✓				
ℓ_4: bakery				✓			✓				✓	
ℓ_5: community center	✓				✓			✓				

to be operated. MMUs are naturally expensive and the number of required vehicles should therefore be considered during tactical MMU operation planning.

In this chapter, we consider two approaches to compute tactical MMU operation plans. The first approach (Section 9.1) builds on the strategic MMU operation plans derived in Phase 1 by partitioning them into tactical MMU operation plans such that each session service provides a similar coverage while minimizing the number of required MMUs. The second approach (Section 9.2) extends the results from Chapter 8 to allow for a combined consideration of Phases 1 and 2 and thus directly yields a tactical MMU operation plan. While this combination of Phases 1 and 2 offers a higher optimization potential, it requires more empirical data which is why we consider both presented approaches as feasible alternatives.

9.1 Partitions of Strategic MMU Operation Plans

In this section, we consider the tactical MMU operation planning based on a strategic MMU operation plan. To that end, let a strategic MMU operation plan $m^{\mathrm{S}}\colon L \to \mathbb{N}$ be given. Recall, that the patient demand origins $v \in V$ used in Phase 1 are aggregated and specify steerable and unsteerable patient demands at a weekly basis. As a result, they provide no information regarding the patient demand in a particular session of the week. Assuming that patient demands are evenly distributed among all sessions, we aim at a tactical MMU operation plan such that the session services minimize the patients' access distances to their closest treatment facility on each session of the week. To obtain such a tactical MMU operation plan, we partition the given strategic MMU operation plan m^{S}.

Definition 9.2. A tactical MMU operation plan $m^{\mathrm{T}}\colon \boldsymbol{L} \to \{0,1\}$ is called a *partition* of the strategic MMU operation plan $m^{\mathrm{S}}\colon L \to \mathbb{N}$ if and only if $m^{\mathrm{S}}_{\ell} = \sum_{\lambda\in\Lambda} m^{\mathrm{T}}_{(\ell,\lambda)}$ for all $\ell \in L$. By definition, all partitions m^{T} of m^{S} have cost $c(m^{\mathrm{T}}) = c(m^{\mathrm{S}})$. We call a partition m^{T} of m^{S} *minimal*, if it requires a minimum number of MMUs for its operation, i.e., $\nu(m^{\mathrm{T}}) \leq \nu(\overline{m}^{\mathrm{T}})$ for all partitions $\overline{m}^{\mathrm{T}}$ of m^{S}.

We assume in the following that $m^{\mathrm{S}}_{\ell} \leq |\Lambda|$ for all $\ell \in L$ as we forbid parallel MMU operations and want to ensure the existence of a partition. Moreover, as each MMU can service at

most $|\Lambda|$ session per week, each minimal partition m^{T} of m^{S} requires exactly $\nu(m^{\mathrm{T}}) = \lceil \sum_{\ell \in L} m^{\mathrm{S}}_{\ell} / |\Lambda| \rceil =: \nu(m^{\mathrm{S}})$ MMUs by evenly distributing services among vehicles.

Next to the MMU operations, the access distances of the patient demand origins in each session depend on the availability of the regular practices P. To that end, let $o\colon \boldsymbol{P} \to \{0,1\}$ indicate the practices' opening hours as only the practices in $P_\lambda(o) \coloneqq \{p \in P : o(p,\lambda) = 1\}$ are available in session $\lambda \in \Lambda$. Analogous to session services, we abbreviate the available practices in session $\lambda \in \Lambda$ by P_λ to ease notation. We can now define the *covering radius* of a session service $L_\lambda \subseteq L$ as $r(L_\lambda) \coloneqq \max_{v \in V} \min_{k \in L_\lambda \cup P_\lambda} \operatorname{dist}(v,k)$. As a result, the covering radius measures the maximum access distance of the demand origins in session $\lambda \in \Lambda$ if session service L_λ is operated. Analogously, we define the *covering radius* of a tactical MMU operation plan m^{T} as the maximum covering radius among all its session services, i.e., $r(m^{\mathrm{T}}) \coloneqq \max_{\lambda \in \Lambda} r(L_\lambda)$. We can now formalize the tactical partitioning problem for MMUs.

Definition 9.3. (TPMMU) Let $m^{\mathrm{S}}\colon L \to \mathbb{N}$ be a strategic MMU operation plan, Λ the sessions of the week, V the patient demand origins, and P the practices with opening hours $o\colon P \times \Lambda \to \{0,1\}$. Then the *tactical partitioning problem for MMUs* (TPMMU) asks for a minimal partition m^{T} of m^{S} with minimum covering radius.

The TPMMU is a partitioning variant of the k-center problem and we investigate its complexity before considering solution approaches. To that end, we begin by introducing a variant of the dominating set problem which asks for a dominating set containing half of a graph's nodes.

Definition 9.4. Let $G = (V(G), E(G))$ be a graph with an even number of nodes, i.e., $V(G) = \{w_1, \dots, w_n\}$ where $n \mod 2 = 0$. Then the $\frac{1}{2}$*-dominating set problem* asks whether there exits $D \subseteq V(G)$ with $|D| = \frac{n}{2}$ such that for all $w_i \in V(G) \setminus D$ there exists $w_j \in D$ with $\{w_i, w_j\} \in E(G)$.

We show that this specialization of the dominating set problem is strongly $\mathcal{NP}$-complete.

Theorem 9.5. *The $\frac{1}{2}$-dominating set problem is strongly $\mathcal{NP}$-complete.*

Proof. We prove the strong $\mathcal{NP}$-completeness of the $\frac{1}{2}$-dominating set problem by a reduction from the (general) dominating set problem which is known to be strongly $\mathcal{NP}$-complete (Garey and Johnson, 1979). Recall that in the dominating set problem, we are given a graph $G = (V(G), E(G))$ with node set $V(G) = \{w_1, \dots, w_n\}$ and an integer $k \leq n$ and have to decide whether there exists $D \subseteq V(G)$ with $|D| = k$ such that for all $w_i \in V(G) \setminus D$ there exists $w_j \in D$ with $\{w_i, w_j\} \in E(G)$.

Given an instance $\mathcal{I} = (G, k)$ of the dominating set problem with $V(G) = \{w_1, \dots, w_n\}$, we construct an instance $\mathcal{I}' = (G')$ of the $\frac{1}{2}$-dominating set problem with $V(G') = \{w'_1, \dots, w'_{n'}\}$ as follows. If $k = \frac{n}{2}$, $\mathcal{I}$ is an instance of the $\frac{1}{2}$-dominating set problem and the reduction

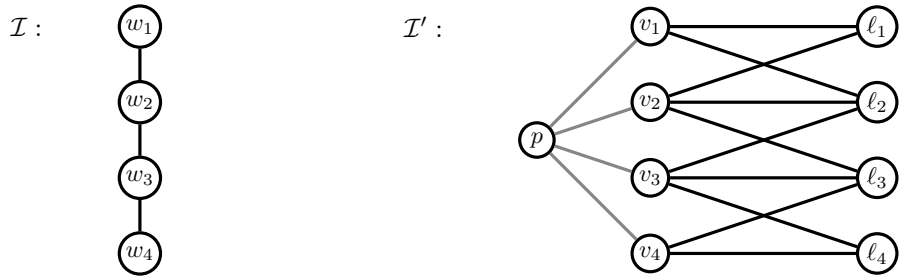


Fig. 9.1.: Constructed TPMMU instance $\mathcal{I}'$ for given $\frac{1}{2}$-dominating set instance $\mathcal{I} = (G)$. Edges in $\mathcal{I}'$ indicate distances of length 0. Dominating set $D = \{w_2, w_4\}$ corresponds to minimal partition m^{T} with session service $L_{\lambda_1} = \{\ell_2, \ell_4\}$ and covering radius $r(m^{\mathrm{T}}) = r(L_{\lambda_1}) = 0$.

is trivial. Otherwise, we add singletons or a star to G to adjust the ratio between n and k. Specifically, if $k < \frac{n}{2}$, we choose $G' = G \cup H$ where H consists of $n - 2k$ singletons, i.e., $V(H) = \{v_1, \ldots, v_{n-2k}\}$ and $E(H) = \emptyset$. By construction, $n' = 2n - 2k$ which is by assumption positive and even. If $k > \frac{n}{2}$, we choose $G' = G \cup H$ where H is a star with $2(k+1) - n$ nodes, i.e., $V(H) = \{v_1\} \cup \{v_2, \ldots, v_{2(k+1)-n}\}$ and $E(H) = \{\{v_1, v_i\} : 2 \leq i \leq 2(k+1) - n\}$. Again, we have by construction that $n' = 2(k+1)$ which is positive and even.

We show that $\mathcal{I}$ is a Yes-instance if and only if $\mathcal{I}'$ is a Yes-instance. Assume that $\mathcal{I}$ is a Yes-instance and let $D \subseteq V(G)$ be a dominating set in G of cardinality $|D| = k$. To show that there exists a dominating set $D' \subseteq V(G')$ in G' with $|D'| = \frac{n'}{2}$, we distinguish cases. If $k < \frac{n}{2}$, we add the singletons in H to the dominating set, i.e., we choose $D' = D \cup V(H)$. Then obviously D' is a dominating set in G' with cardinality $|D'| = |D| + n - 2k = n - k = \frac{n'}{2}$. If $k > \frac{n}{2}$, we add the center of the star H to the dominating set, i.e., we set $D' = D \cup \{v_1\}$. By construction, D' is a dominating set in G' with cardinality $|D'| = |D| + 1 = k + 1 = \frac{n'}{2}$. The converse direction can be shown analogously.

Last but not least, we note that the $\frac{1}{2}$-dominating set problem is a special case of the dominating set problem and therefore obviously in $\mathcal{NP}$. □

We can now show that the TPMMU with just two sessions per week is $\mathcal{NP}$-hard.

Theorem 9.6. *The TPMMU with two sessions per week, i.e., $|\Lambda| = 2$, is strongly $\mathcal{NP}$-hard.*

Proof. We show the strong $\mathcal{NP}$-hardness of the TPMMU by a reduction from the $\frac{1}{2}$-dominating set problem. Given an instance $\mathcal{I} = (G)$ of the $\frac{1}{2}$-dominating set problem with $V(G) = \{w_1, \ldots, w_n\}$, we construct an instance $\mathcal{I}'$ of the TPMMU as follows. We choose the patient demand origins $V = \{v_1, \ldots, v_n\}$ that represents the nodes of G and the MMU operation sites $L = \{\ell_1, \ldots, \ell_n\}$ which are going to encode the dominating set. Furthermore let $\Lambda = \{\lambda_1, \lambda_2\}$, $m_\ell^{\mathsf{S}} = 1$ for all $\ell \in L$, and $P = \{p\}$ with $o(p, \lambda_1) = 0$ and $o(p, \lambda_2) = 1$. Finally, we define the distances between all $v_i \in V$ and $\ell_j \in L$ as

$$\mathrm{dist}(v_i, \ell_j) = \begin{cases} 0 & \text{if } i = j \vee \{w_i, w_j\} \in E(G) \\ 1 & \text{else} \end{cases}$$

and set $\mathrm{dist}(v_i, p) = 0$ for all $v_i \in V$. By construction, every minimal partition m^{T} of m^{S} requires exactly $\nu(m^{\mathrm{S}}) = \lceil |L|/2 \rceil = \frac{n}{2}$ MMUs and for both sessions $\lambda \in \Lambda$ it must hold that $|L_\lambda| = \frac{n}{2}$. Moreover, note that for every partition m^{T} of m^{S} we have that $r(L_{\lambda_2}) = 0$ due to the open practice $p \in P$ which implies that $r(m^{\mathrm{T}}) = r(L_{\lambda_1})$. The construction of $\mathcal{I}'$ for an exemplary $\frac{1}{2}$-dominating set instance is illustrated in Figure 9.1.

We show that $\mathcal{I}$ is a Yes-instance if and only if $\mathcal{I}'$ has a minimal partition m^{T} with covering radius $r(m^{\mathrm{T}}) = r(L_{\lambda_1}) = 0$. Assume $\mathcal{I}$ is a Yes-instance and let $D \subseteq V(G)$ be a dominating set of size $|D| = \frac{n}{2}$. Then, for the minimal partition m^{T} of m^{S} with $L_{\lambda_1} = \{\ell_i \in L : w_i \in D\}$ and $L_{\lambda_2} = L \setminus L_{\lambda_1}$ we have that $r(m^{\mathrm{T}}) = r(L_{\lambda_1}) = 0$ as D is a dominating set. The converse direction can be shown analogously. □

The $\frac{1}{2}$-dominating set problem can be naturally extended to general $k \in \mathbb{N} \setminus \{0\}$, i.e., the $\frac{1}{k}$-*dominating set problem* on G that asks for a dominating set $D \subseteq V(G)$ of cardinality $|D| = \frac{|V(G)|}{k}$. Generalizing the reduction from Theorem 9.5, it follows that for all $k \geq 2$ the $\frac{1}{k}$-dominating set problem is strongly $\mathcal{NP}$-complete. Thus, the construction in the reduction above yields the following.

Corollary 9.7. *For each $k \in \mathbb{N}$ with $k \geq 2$, the TPMMU with $|\Lambda| = k$ is strongly $\mathcal{NP}$-hard.*

Moreover, due to our choice of distances in the previous reduction, any α-approximation algorithm for the TPMMU could decide the $\frac{1}{k}$-dominating set problem in polynomial time which yields the following inapproximability result.

Corollary 9.8. *For each $k \in \mathbb{N}$ with $k \geq 2$, it is $\mathcal{NP}$-hard to approximate the TPMMU with $|\Lambda| = k$ within a constant approximation factor $\alpha > 1$.*

The distances in the proof of Theorem 9.6 do not satisfy the triangle inequality. As a result, both Corollaries 9.7 and 9.8 do not hold for the *metric* TPMMU in which the triangle inequality must hold. By adjusting the distance function, we can show that also the metric TPMMU is $\mathcal{NP}$-hard and obtain a slightly weaker inapproximability result.

Theorem 9.9. *For each $k \in \mathbb{N}$ with $k \geq 2$, the metric TPMMU with $|\Lambda| = k$ is strongly $\mathcal{NP}$-hard.*

Proof. To show the strong $\mathcal{NP}$-hardness of the metric TPMMU, we use the reduction from the proof of Theorem 9.6 and adjust the distances between the demand origins $v_i \in V$ and operation sites $\ell_j \in L$ to

$$\mathrm{dist}(v_i, \ell_j) = \begin{cases} 1 & \text{if } i = j \vee \{w_i, w_j\} \in E(G) \\ 2 & \text{else} \end{cases}$$

and set $\mathrm{dist}(v_i, p) = 1$ for all $v_i \in V$. The adjusted distances obviously satisfy the triangle inequality. Moreover, analogous to Theorem 9.6, it holds that a $\frac{1}{2}$-dominating set instance

$\mathcal{I}$ is a Yes-instace if and only if the corresponding metric TPMMU instance $\mathcal{I}'$ with adjusted distances has a minimal partition m^{T} with covering radius $r(m^{\mathrm{T}}) = 1$. Last but not least, we note that this construction generalizes to the $\frac{1}{k}$-dominating set problem for $k \geq 2$. □

With the adjusted distances in the proof of Theorem 9.9, the covering radius of a minimal partition is either one or two. As a result, any α-approximation algorithm for the metric TPMMU with $\alpha < 2$ could decide the $\frac{1}{k}$-dominating set problem in polynomial time.

Corollary 9.10. *For each $k \in \mathbb{N}$ with $k \geq 2$, it is $\mathcal{NP}$-hard to approximate the metric TPMMU with $|\Lambda| = k$ within a constant approximation factor $1 < \alpha < 2$.*

Hence, while we have shown that there can be no α-approximation algorithm with constant approximation factor for the general TPMMU (unless $\mathcal{P} = \mathcal{NP}$), there can be a 2-approximation algorithm for the TPMMU if the distances satisfy the triangle inequality.

Having established that solving the TPMMU is non-trivial if a week has at least two sessions, we devise an exact solution approach based on a compact integer linear program. To that end, let variables $x_{\boldsymbol{\ell}} \in \{0,1\}$ for $\boldsymbol{\ell} = (\ell, \lambda) \in \boldsymbol{L}$ denote whether site $\ell \in L$ is serviced in session $\lambda \in \Lambda$. To model the access distances of the patient demand origins, let variables $z_{v\boldsymbol{k}} \in \{0,1\}$ for $v \in V$ and $\boldsymbol{k} = (k, \lambda) \in \boldsymbol{L} \cup \boldsymbol{P}$ denote whether $k \in L \cup P$ is the closest treatment facility to $v \in V$ in session $\lambda \in \Lambda$. Finally, we require an auxiliary variable $\Omega \geq 0$ to linearize the outer maximization in the computation of the covering radius. We can now formulate the TPMMU as follows:

$$
\begin{aligned}
(\mathrm{TP}) \quad & \min_{x,\, z,\, \Omega} && \Omega && && (9.1a)\\
& \text{s.t.} && \Omega \geq \operatorname{dist}(v,k)\, z_{v\boldsymbol{k}} && \forall v \in V,\, \forall \boldsymbol{k} = (k,\lambda) \in \boldsymbol{L} \cup \boldsymbol{P} && (9.1b)\\
& && \sum_{k \in L \cup P_\lambda} z_{v(k,\lambda)} \geq 1 && \forall v \in V,\, \forall \lambda \in \Lambda && (9.1c)\\
& && z_{v\boldsymbol{\ell}} \leq x_{\boldsymbol{\ell}} && \forall v \in V,\, \forall \boldsymbol{\ell} \in \boldsymbol{L} && (9.1d)\\
& && \sum_{\lambda \in \Lambda} x_{(\ell,\lambda)} = m^{\mathrm{S}}_{\ell} && \forall \ell \in L && (9.1e)\\
& && \sum_{\ell \in L} x_{(\ell,\lambda)} \leq \nu(m^{\mathrm{S}}) && \forall \lambda \in \Lambda && (9.1f)\\
& && x_{\boldsymbol{\ell}} \in \{0,1\} && \forall \boldsymbol{\ell} \in \boldsymbol{L} && (9.1g)\\
& && z_{v\boldsymbol{k}} \in \{0,1\} && \forall v \in V,\, \forall \boldsymbol{k} \in \boldsymbol{L} \cup \boldsymbol{P}. && (9.1h)
\end{aligned}
$$

In this formulation, inequalities (9.1b) ensure that the covering radius is correctly modeled, inequalities (9.1c) and (9.1d) model the assignment of the demand origins to their closest treatment facility in each session, and constraints (9.1e) and (9.1f) require the tactical MMU operation plan encoded by x to be a minimal partition of m^{S}. Formulation (TP) is a compact integer linear formulation for the TPMMU. Given an optimal solution (x, z, Ω) to (TP), an optimal solution to the TPMMU is given by the minimal partition $m^{\mathrm{T}}_{\boldsymbol{\ell}} = x_{\boldsymbol{\ell}}$ for all $\boldsymbol{\ell} \in \boldsymbol{L}$.

Remark, that our definition of the covering radius can be dominated by sessions in which many primary care physicians are closed. An example for such a setting can be observed in the computational study in Chapter 11, as most practices in Germany are closed on Wednesday afternoons. In such cases, we want to put weight on each session's covering radius by considering the sum of the covering radii, i.e., $r(m^{\mathrm{T}}) \coloneqq \sum_{\lambda \in \Lambda} r(L_\lambda)$. Formulation (TP) can be easily adopted to this alternative objective function and we note, that the (metric) TPMMU minimizing the sum of the covering radii remains strongly $\mathcal{NP}$-hard.

In the following section, we consider the combined strategic and tactical planning for MMUs. While a combined consideration of Phases 1 and 2 can lead to better solutions, it comes with the downside of requiring more empirical data and being computationally more challenging.

9.2 Combined Strategic and Tactical Planning for MMUs

Chapter 8 considered the (robust) strategic planning problem for MMUs in a session-aggregated form. That is, we modeled the patient demands at each demand origin through a single aggregated value and decided on the total weekly number of MMU sessions at each MMU operation site. Such an aggregation has several shortcomings, as it artificially smoothes out patient demands and entails the previously discussed post-processing step to partition strategic MMU operation plans into tactical MMU operation plans.

To overcome these drawbacks, we can disaggregate the strategic planning problem for MMUs by considering session-specific demands, treatment capacities, and MMU operations. Thereby, steerable patient demands are allowed to be assigned between sessions to balance out each session's workload. The combined strategic tactical planning problem for MMUs (STMMU) then asks for a tactical MMU operation plan $m^{\mathrm{T}} \colon \boldsymbol{L} \to \{0, 1\}$ that satisfies all patient demands at minimum cost.

We formalize this problem by considering the session-specific treatment capacity $\bar{b}_{\boldsymbol{p}} \in \mathbb{N}$ for every session-expanded practice $\boldsymbol{p} \in \boldsymbol{P}$. To model session-specific patient demands, we consider the session-expanded demand origins $\boldsymbol{V} \coloneqq V \times \Lambda$ with a steerable patient demand $d_{\boldsymbol{v}} \in \mathbb{N}$ and an unsteerable patient demand $u_{\boldsymbol{v}} \in \mathbb{N}$ for each $\boldsymbol{v} \in \boldsymbol{V}$. While unsteerable patient demands immediately visit the closest considered operating treatment facility, steerable patient demands can be shifted between sessions. Thus, we model two independent consideration sets for each $\boldsymbol{v} = (v, \lambda) \in \boldsymbol{V}$: a consideration set $\boldsymbol{N}^d(\boldsymbol{v}) \subseteq \boldsymbol{L} \cup \boldsymbol{P}$ for the steerable patient demands, and a consideration set $\boldsymbol{N}^u(\boldsymbol{v}) \subseteq (L \cup P) \times \{\lambda\}$ for the unsteerable patient demands. As a result, we have to extend the definition of an assignment of the steerable patient demands (Definition 8.2).

Definition 9.11. A *session-specific assignment* of the steerable patient demands is a set of functions $\{f_{\boldsymbol{v}}\}_{\boldsymbol{v}\in\boldsymbol{V}}$ with $f_{\boldsymbol{v}}\colon \boldsymbol{N}^d(\boldsymbol{v}) \to \mathbb{N}$ that distribute all steerable patient demands within their respective session-expanded consideration set, i.e., $\sum_{\boldsymbol{k}\in\boldsymbol{N}^d(\boldsymbol{v})} f_{\boldsymbol{v}}(\boldsymbol{k}) = d_{\boldsymbol{v}}$ for all $\boldsymbol{v} \in \boldsymbol{V}$.

Next, we define *feasible* tactical MMU operation plans. To ease notation, let $\boldsymbol{N}^d(\boldsymbol{k}) \coloneqq \{\boldsymbol{v} \in \boldsymbol{V} : \boldsymbol{k} \in \boldsymbol{N}^d(\boldsymbol{v})\}$ ($\boldsymbol{N}^u(\boldsymbol{k}) \coloneqq \{\boldsymbol{v} \in \boldsymbol{V} : \boldsymbol{k} \in \boldsymbol{N}^u(\boldsymbol{v})\}$) denote all session-expanded patient demand origins whose (un-)steerable patient demands can target the treatment facility $\boldsymbol{k} \in \boldsymbol{L} \cup \boldsymbol{P}$. Moreover, let $\boldsymbol{k}^{\min}_{\boldsymbol{v}}(m^{\mathrm{T}}) \in \boldsymbol{N}^u(\boldsymbol{v})$ denote the closest considered operating treatment facility which is targeted by all unsteerable patient demands originating in $\boldsymbol{v} \in \boldsymbol{V}$ for given tactical MMU operation plan m^{T}.

Definition 9.12. A tactical MMU operation plan m^{T} is *feasible* if there exists a session-specific assignment of the steerable patient demands $\{f_{\boldsymbol{v}}\}_{\boldsymbol{v}\in\boldsymbol{V}}$ that respects the session-specific treatment capacity at each treatment facility $\boldsymbol{k} \in \boldsymbol{L} \cup \boldsymbol{P}$, that is

$$\sum_{\boldsymbol{v}\in\boldsymbol{V}:\boldsymbol{k}^{\min}_{\boldsymbol{v}}(m^{\mathrm{T}})=\boldsymbol{k}} u_{\boldsymbol{v}} + \sum_{\boldsymbol{v}\in\boldsymbol{N}^d(\boldsymbol{k})} f_{\boldsymbol{v}}(\boldsymbol{k}) \leq \begin{cases} \bar{b}_{\boldsymbol{k}} & \text{if } \boldsymbol{k} \in \boldsymbol{P}, \\ \hat{b}\, m^{\mathrm{T}}_{\boldsymbol{k}} & \text{if } \boldsymbol{k} \in \boldsymbol{L}. \end{cases}$$

Finally, we can employ the notion of a feasible tactical MMU operation plan to formalize the definition of the combined strategic tactical planning problem for MMUs.

Definition 9.13 (STMMU)**.** Let Λ denote the sessions of the week and let $\ell \in L$ be the potential MMU operation sites with setup costs $c_\ell \in \mathbb{N}$. Moreover, let $p \in P$ be the existing practices with treatment capacities $\bar{b}_{(p,\lambda)} \in \mathbb{N}$ in session $\lambda \in \Lambda$. In every session $\lambda \in \Lambda$, each patient demand origin $v \in V$ has steerable and unsteerable demands $d_{(v,\lambda)}, u_{(v,\lambda)} \in \mathbb{N}$ that can be serviced within the consideration sets $\boldsymbol{N}^d((v,\lambda)) \subseteq (L \cup P) \times \Lambda$ and $\boldsymbol{N}^u((v,\lambda)) \subseteq (L \cup P) \times \{\lambda\}$, respectively. Then, the *combined strategic tactical planning problem for MMUs* (STMMU) asks for a feasible tactical MMU operation plan of minimum cost, where every operated MMU session induces the cost $\hat{c} \in \mathbb{N}$ and yields a treatment capacity $\hat{b} \in \mathbb{N}$.

As the STMMU only allows for a single MMU operation at every site $\boldsymbol{\ell} \in \boldsymbol{L}$, the STMMU does not generalize the SMMU and thus the problem's strong $\mathcal{NP}$-hardness does not follow from Theorem 8.5. However, the reduction referenced in the proof of Theorem 8.5 is still applicable for the STMMU, as all subsets in this reduction are chosen at most once.

Corollary 9.14. *The STMMU is strongly $\mathcal{NP}$-hard.*

Comparing the STMMU to the SMMU, we can observe that both problems are closely related. In the following, we devise an integer linear programming formulation for the STMMU which is nearly identical to formulation (Det) from Section 8.1 and emphasizes the problems' common structure. Let variables $y_\ell \in \{0,1\}$ indicate whether site $\ell \in L$ is set up, let variables $x_{\boldsymbol{\ell}} \in \{0,1\}$ decide whether site $\boldsymbol{\ell} \in \boldsymbol{L}$ is serviced by an MMU, and let variables $z_{\boldsymbol{vk}} \in \mathbb{N}$

determine the steerable demand originating in $\boldsymbol{v} \in \boldsymbol{V}$ that is assigned to treatment facility $\boldsymbol{k} \in \boldsymbol{N}^d(\boldsymbol{v})$. Moreover, let variables $w_{\boldsymbol{vk}} \in \{0,1\}$ indicate the closest operating treatment facility $\boldsymbol{k} \in \boldsymbol{N}^u(\boldsymbol{v})$ that is targeted by all unsteerable demands originating in $\boldsymbol{v} \in \boldsymbol{V}$. To that end, let $\boldsymbol{\pi_v} \colon \{1, \dots, |\boldsymbol{N}^u(\boldsymbol{v})|\} \to \boldsymbol{N}^u(\boldsymbol{v})$ define an order on the consideration set $\boldsymbol{N}^u(\boldsymbol{v})$ that is non-decreasing with respect to the treatment facility's distance $\mathrm{dist} \colon \boldsymbol{V} \times (\boldsymbol{L} \cup \boldsymbol{P}) \to \mathbb{N}$ to demand origin $\boldsymbol{v} \in \boldsymbol{V}$. As in Section 8.1, we denote all MMU operation sites and practices within the consideration set of unsteerable demands at $\boldsymbol{v} \in \boldsymbol{V}$ by $\boldsymbol{N}^u_{\boldsymbol{L}}(\boldsymbol{v}) \coloneqq \boldsymbol{N}^u(\boldsymbol{v}) \cap \boldsymbol{L}$ and $\boldsymbol{N}^u_{\boldsymbol{P}}(\boldsymbol{v}) \coloneqq \boldsymbol{N}^u(\boldsymbol{v}) \cap \boldsymbol{P}$, respectively. We can now formulate the STMMU as follows:

$$
\begin{aligned}
(\lambda\mathrm{Det}) \quad \min_{y,\, x,\, z,\, w} \quad & \sum_{\ell \in L} c_\ell \, y_\ell + \sum_{\boldsymbol{\ell} \in \boldsymbol{L}} \hat{c} \, x_{\boldsymbol{\ell}} && && \text{(9.2a)} \\
\text{s.t.} \quad & x_{\boldsymbol{\ell}} \leq y_\ell && \forall \boldsymbol{\ell} = (\ell, \lambda) \in \boldsymbol{L} && \text{(9.2b)} \\
& \sum_{\boldsymbol{k} \in \boldsymbol{N}^d(\boldsymbol{v})} z_{\boldsymbol{vk}} \geq d_{\boldsymbol{v}} && \forall \boldsymbol{v} \in \boldsymbol{V} && \text{(9.2c)} \\
& \sum_{\boldsymbol{v} \in \boldsymbol{N}^d(\boldsymbol{\ell})} z_{\boldsymbol{v\ell}} + \sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{\ell})} u_{\boldsymbol{v}} \, w_{\boldsymbol{v\ell}} \leq \hat{b} \, x_{\boldsymbol{\ell}} && \forall \boldsymbol{\ell} \in \boldsymbol{L} && \text{(9.2d)} \\
& \sum_{\boldsymbol{v} \in \boldsymbol{N}^d(\boldsymbol{p})} z_{\boldsymbol{vp}} + \sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{p})} u_{\boldsymbol{v}} \, w_{\boldsymbol{vp}} \leq \bar{b}_{\boldsymbol{p}} && \forall \boldsymbol{p} \in \boldsymbol{P} && \text{(9.2e)} \\
& \sum_{\boldsymbol{k} \in \boldsymbol{N}^u(\boldsymbol{v})} w_{\boldsymbol{vk}} \geq 1 && \forall \boldsymbol{v} \in \boldsymbol{V} && \text{(9.2f)} \\
& w_{\boldsymbol{v\ell}} \leq x_{\boldsymbol{\ell}} && \forall \boldsymbol{v} \in \boldsymbol{V},\, \forall \boldsymbol{\ell} \in \boldsymbol{N}^u_{\boldsymbol{L}}(\boldsymbol{v}) && \text{(9.2g)} \\
& w_{\boldsymbol{v\ell}} \geq x_{\boldsymbol{\ell}} - \sum_{i=1}^{\boldsymbol{\pi_v}^{-1}(\boldsymbol{\ell})-1} w_{\boldsymbol{v}, \boldsymbol{\pi_v}(i)} && \forall \boldsymbol{v} \in \boldsymbol{V},\, \forall \boldsymbol{\ell} \in \boldsymbol{N}^u_{\boldsymbol{L}}(\boldsymbol{v}) && \text{(9.2h)} \\
& w_{\boldsymbol{vp}} \geq 1 - \sum_{i=1}^{\boldsymbol{\pi_v}^{-1}(\boldsymbol{p})-1} w_{\boldsymbol{v}, \boldsymbol{\pi_v}(i)} && \forall \boldsymbol{v} \in \boldsymbol{V},\, \forall \boldsymbol{p} \in \boldsymbol{N}^u_{\boldsymbol{P}}(\boldsymbol{v}) && \text{(9.2i)} \\
& x_{\boldsymbol{\ell}} \in \{0,1\},\ y_\ell \in \{0,1\} && \forall \boldsymbol{\ell} = (\ell, \lambda) \in \boldsymbol{L} && \text{(9.2j)} \\
& w_{\boldsymbol{vk}} \in \{0,1\} && \forall \boldsymbol{v} \in \boldsymbol{V},\, \forall \boldsymbol{k} \in \boldsymbol{N}^u(\boldsymbol{v}) && \text{(9.2k)} \\
& z_{\boldsymbol{vk}} \in \mathbb{N} && \forall \boldsymbol{v} \in \boldsymbol{V},\, \forall \boldsymbol{k} \in \boldsymbol{N}^d(\boldsymbol{v}). && \text{(9.2l)}
\end{aligned}
$$

Putting formulations (Det) and $(\lambda\mathrm{Det})$ side-by-side, we can confirm that the disaggreation of sessions leads to a structurally identical problem. Consequently, all results from Section 8.1 can be directly transferred to STMMU. In particular, we can analogously show the correctness of the formulation $(\lambda\mathrm{Det})$.

Theorem 9.15. $(\lambda\mathrm{Det})$ *is an integer linear formulation for the STMMU.*

Moreover, we can apply the Benders decomposition approach from Section 8.1 to $(\lambda\mathrm{Det})$ to obtain the following analogous result.

Theorem 9.16. *Constraints* (9.2c)–(9.2e), (9.2l) *in* $(\lambda\mathrm{Det})$ *can be equivalently substituted by*

$$\sum_{\boldsymbol{v}\in\boldsymbol{U}} d_{\boldsymbol{v}} + \sum_{\boldsymbol{k}\in\boldsymbol{N}^d(\boldsymbol{U})} \sum_{\boldsymbol{v}\in\boldsymbol{N}^u(\boldsymbol{k})} u_{\boldsymbol{v}}\, w_{\boldsymbol{vk}} \leq \sum_{\boldsymbol{\ell}\in\boldsymbol{N}^d_{\boldsymbol{L}}(\boldsymbol{U})} \hat{b}\, x_{\boldsymbol{\ell}} + \sum_{\boldsymbol{p}\in\boldsymbol{N}^d_{\boldsymbol{P}}(\boldsymbol{U})} \bar{b}_{\boldsymbol{p}} \qquad \forall \boldsymbol{U}\subseteq\boldsymbol{V}, \tag{9.3}$$

where $\boldsymbol{N}^d(\boldsymbol{U}) \coloneqq \bigcup_{\boldsymbol{v}\in\boldsymbol{U}} \boldsymbol{N}^d(\boldsymbol{v})$, $\boldsymbol{N}^d_{\boldsymbol{L}}(\boldsymbol{U}) \coloneqq \boldsymbol{N}^d(\boldsymbol{U}) \cap \boldsymbol{L}$, *and* $\boldsymbol{N}^d_{\boldsymbol{P}}(\boldsymbol{U}) \coloneqq \boldsymbol{N}^d(\boldsymbol{U}) \cap \boldsymbol{P}$ *for* $\boldsymbol{U}\subseteq\boldsymbol{V}$.

The resulting Benders reformulation of $(\lambda\mathrm{Det})$ will be denoted by $(\lambda\mathrm{Det\text{-}B})$. Obviously, we can also transfer all results from Section 8.2 to the STMMU to obtain a constraint generation procedure for the robust combined strategic tactical planning problem for MMUs with interval and budgeted uncertainty sets.

The integration of the required number of MMUs into the STMMU is straightforward. However, instead of fixing the number of vehicles beforehand as it was done in the TPMMU, we assume a fixed cost $\bar{c}\in\mathbb{N}$ per MMU and add the resulting cost term $\bar{c}\cdot\nu(m^{\mathrm{T}})$ to the total cost of the tactical MMU operation plan m^{T}. We can implement this extension in formulations $(\lambda\mathrm{Det})$ and $(\lambda\mathrm{Det\text{-}B})$ by adding a non-negative variable $\nu\geq 0$ which models the number of MMUs required to operate the tactical MMU operation plan encoded by x via the constraints

$$\nu \geq \sum_{\ell\in L} x_{(\ell,\lambda)} \qquad \forall\lambda\in\Lambda. \tag{9.4}$$

The use of MMUs can then be penalized by adding the additional cost term $\bar{c}\cdot\nu$ to the objective (9.2a).

Concluding this chapter, recall that we did allow for at most one MMU operation at each session-expanded location $\boldsymbol{\ell}\in\boldsymbol{L}$ in all our approaches. This restriction can be dropped, as all result of this chapter generalize to parallel MMU operations. However, this requires additional constraints that complicate our models while providing minimal gain since parallel MMU operations rarely occur in the considered low density population setting. Even worse, in the TPMMU we can observe that parallel MMU operations provide no value at all as they do not affect the covering radius. We will therefore not elaborate on this model extension in this thesis.

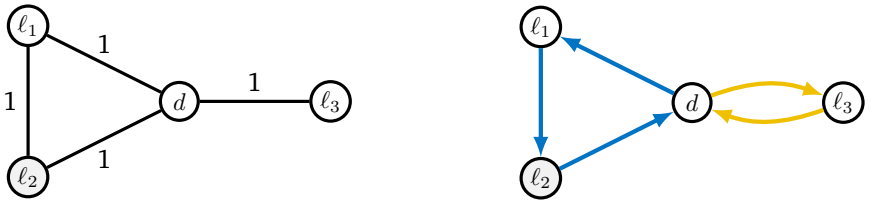
10 Phase 3: Vehicle Routing for MMUs

Phase 3 in P3MMU considers the vehicle routing for a given tactical MMU operation plan $m^{\mathrm{T}}\colon \boldsymbol{L} \to \{0,1\}$ as it results from the tactical planning phase that we studied in the previous chapter. Recall, that we consider the mode of operation in which MMUs are stationed at a fixed home depot to which they return at the end of each day. Consequently, every vehicle's daily route is uniquely defined by the MMU operation sites that are serviced in the morning and afternoon session. Moreover, if the assignment of vehicles to depots is given (which is particularly the case in the one depot setting), the routing problems for each day of the week become independent as all vehicles return to their assigned depot at the end of each day. Hence, we can decompose the vehicle routing problem for the entire week into several daily vehicle routing problems that we investigate in the following. We thereby differentiate whether there is only a single or several depots.

10.1 MMU Routing with a Single Depot

In this section, we investigate the vehicle routing problem for MMUs with a single depot (VRMMU). To that end, let d denote the single depot at which all MMUs are stationed and let $(L^{\mathrm{AM}}, L^{\mathrm{PM}})$ with $L^{\mathrm{AM}} \subseteq L$, $L^{\mathrm{PM}} \subseteq L$ be a pair of MMU operation sites which need to be serviced in the respective morning and afternoon session of the same day. We assume throughout this chapter that w.l.o.g. $|L^{\mathrm{AM}}| \geq |L^{\mathrm{PM}}|$ to avoid the distinction of several analogous cases. As we consider only a single depot, every pair $(\ell_1, \ell_2) \in L^{\mathrm{AM}} \times L^{\mathrm{PM}}$ uniquely defines the vehicle route $d \to \ell_1 \to \ell_2 \to d$ of length $c(\ell_1, \ell_2) \coloneqq \mathrm{dist}(d, \ell_1) + \mathrm{dist}(\ell_1, \ell_2) + \mathrm{dist}(\ell_2, d)$. To indicate that a vehicle route services only one operation site, we introduce the empty site $\emptyset$, i.e., the pair $(\ell, \emptyset)$ with $\ell \in L^{\mathrm{AM}}$ encodes the route $d \to \ell \to d$ of length $c(\ell, \emptyset) \coloneqq \mathrm{dist}(d, \ell) + \mathrm{dist}(\ell, d)$. All vehicle routes can thus be represented by the set $\mathcal{R} \coloneqq \{(\ell_1, \ell_2) : \ell_1 \in L^{\mathrm{AM}} \cup \{\emptyset\}, \ell_2 \in L^{\mathrm{PM}} \cup \{\emptyset\}\}$. To service the operation sites $(L^{\mathrm{AM}}, L^{\mathrm{PM}})$ in the morning and afternoon session of the same day, we require a set of $|L^{\mathrm{AM}}|$ vehicle routes that exhibit the following properties.

Definition 10.1. Let the pair of MMU operation sites $(L^{\mathrm{AM}}, L^{\mathrm{PM}})$ with $L^{\mathrm{AM}} \subseteq L$, $L^{\mathrm{PM}} \subseteq L$ be given. We call a set of vehicle routes $R \subseteq \mathcal{R}$ with $|R| = |L^{\mathrm{AM}}|$ a *route partition* if and only if for all $\ell \in L^{\mathrm{AM}}$ it holds that $|\{(\ell, \ell_2) \in R\}| = 1$ and for all $\ell \in L^{\mathrm{PM}}$ it holds that $|\{(\ell_1, \ell) \in R\}| = 1$. The *cost* of a route partition is defined by the length of its routes, i.e., $c(R) = \sum_{(\ell_1,\ell_2)\in R} c(\ell_1, \ell_2)$.



$L^{\text{AM}} = \{\ell_1, \ell_3\},\ L^{\text{PM}} = \{\ell_2\}$

(a)

$R = \{(\ell_1, \ell_2), (\ell_3, \emptyset)\},\ c(R) = 5$

(b)

Fig. 10.1.: Example of a route partition: (a) VRMMU instance, (b) route partition.

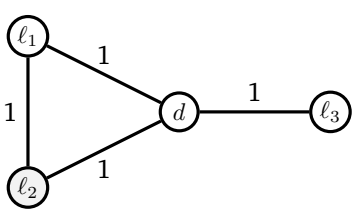
Route partitions ensure that all sites $\ell \in L^{\text{AM}}$ are serviced by exactly one vehicle route in the morning and all sites $\ell \in L^{\text{PM}}$ are services by exactly one vehicle route in the afternoon. An illustration of a route partition can be found in Figure 10.1. By definition, route partitions $R \subseteq \mathcal{R}$ cannot contain routes of the form $(\emptyset, \ell_2) \in \mathcal{R}$ that do not serve a site in the morning session as otherwise $|R| > |L^{\text{AM}}|$. Instead, each route partition must consist of $|L^{\text{PM}}|$ routes that service the morning and the afternoon session sites and $|L^{\text{AM}}| - |L^{\text{PM}}|$ routes that service only the morning session. We can now formally define the vehicle routing problem for MMUs with a single depot.

Definition 10.2 (VRMMU)**.** Given a pair of MMU operation sites $(L^{\text{AM}}, L^{\text{PM}})$ with $L^{\text{AM}} \subseteq L$ and $L^{\text{PM}} \subseteq L$, a vehicle depot d, and distances $\text{dist}\colon L \cup \{d\} \times L \cup \{d\} \to \mathbb{N}$ between them, the *vehicle routing problem for MMUs with a single depot* (VRMMU) asks for a route partition of minimum cost.

To solve the VRMMU, we reduce the problem to a weighted matching problem in a complete bipartite graph $\bar{G} = (V_A \cup V_B, E)$ that we construct as follows. The graph $\bar{G}$ has nodes $V_A \coloneqq \{a_\ell : \ell \in L^{\text{AM}}\}$ and $V_B = V_B^1 \cup V_B^2$ where $V_B^1 \coloneqq \{b_\ell : \ell \in L^{\text{PM}}\}$ and $V_B^2 \coloneqq \{\emptyset_1, \ldots, \emptyset_{|L^{\text{AM}}|-|L^{\text{PM}}|}\}$. The nodes in V_B^2 serve as auxiliary nodes that are needed to encode the $|L^{\text{AM}}| - |L^{\text{PM}}|$ vehicle routes in each route partition that service only the morning session. The set of edges is given by $E = V_A \times V_B$, i.e., $\bar{G}$ is complete bipartite. Each edge in E corresponds to a unique vehicle route which determines the weight of that edge. Specifically, we set $w(\{a_{\ell_1}, b_{\ell_2}\}) = c(\ell_1, \ell_2)$ and $w(\{a_{\ell_1}, \emptyset_i\}) = c(\ell_1, \emptyset)$ for all $a_{\ell_1} \in V_A$, $b_{\ell_2} \in V_B^1$, and $\emptyset_i \in V_B^2$. An example for the construction of the graph $\bar{G}$ can be found in Figure 10.2. By definition, $|V_A| = |V_B| = |L^{\text{AM}}|$ and thus $\bar{G}$ always contains a perfect matching of size $|L^{\text{AM}}|$. We can now show the following.

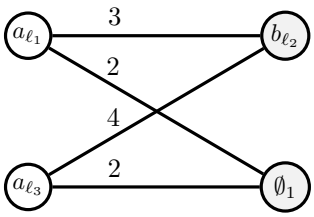
Lemma 10.3. *Every perfect matching $M \subseteq E$ in the graph $\bar{G} = (V_A \cup V_B, E)$ corresponds to a route partition $R \subseteq \mathcal{R}$ with $c(R) = w(M)$ and vice versa.*

Proof. Let $M \subseteq E$ be a perfect matching in the graph $\bar{G} = (V_A \cup V_B, E)$. We construct a route partition $R = R_1 \cup R_2$ with $R_1 \coloneqq \{(\ell_1, \emptyset) \in \mathcal{R} : \{a_{\ell_1}, \emptyset_i\} \in M\}$ and $R_2 \coloneqq \{(\ell_1, \ell_2) \in \mathcal{R} : \{a_{\ell_1}, b_{\ell_2}\} \in M\}$. As M is a perfect matching in $\bar{G}$, it holds that $|R| = |M| = |V_A| = |L^{\text{AM}}|$.



$L^{\text{AM}} = \{\ell_1, \ell_3\},\ L^{\text{PM}} = \{\ell_2\}$

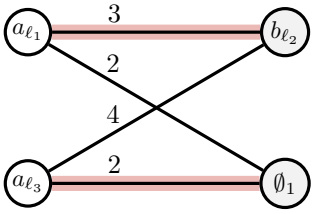
(a)



$\bar{G} = (V_A \cup V_B, E)$

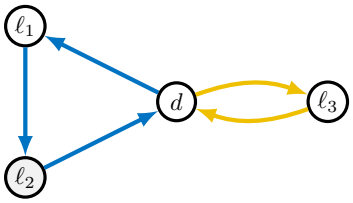
(b)

Fig. 10.2.: Exemplary construction of $\bar{G}$: (a) VRMMU instance where unspecified distances correspond to the length of a shortest path, (b) corresponding graph $\bar{G}$.



$M = \{\{a_{\ell_1}, b_{\ell_2}\}, \{a_{\ell_3}, \emptyset_1\}\},\ w(M) = 5$

(a)



$R = \{(\ell_1, \ell_2), (\ell_3, \emptyset)\},\ c(R) = 5$

(b)

Fig. 10.3.: Correspondence between matching and route partition for VRMMU instance in Figure 10.2: (a) Perfect bipartite matching in $\bar{G}$, (b) corresponding route partition of identical cost.

Moreover, for each $\ell \in L^{\text{AM}}$ we have that $|\{(\ell, \ell_2) \in R\}| = |\delta_M(a_\ell)| = 1$. Analogously, it holds that $|\{(\ell_1, \ell) \in R\}| = |\delta_M(b_\ell)| = 1$ for all $\ell \in L^{\text{PM}}$. As a result, R is indeed a route partition with cost

$$\begin{aligned} c(R) &= c(R_1) + c(R_2) \\ &= \sum_{(\ell_1,\emptyset)\in R_1} c(\ell_1, \emptyset) + \sum_{(\ell_1,\ell_2)\in R_2} c(\ell_1, \ell_2) \\ &= \sum_{\{a_{\ell_1},\emptyset_i\}\in M} w(\{a_{\ell_1}, \emptyset_i\}) + \sum_{\{a_{\ell_1},b_{\ell_2}\}\in M} w(\{a_{\ell_1}, b_{\ell_2}\}) \\ &= w(M). \end{aligned}$$

The converse direction can be shown analogously. □

The correspondence between perfect matchings in the graph $\bar{G}$ and route partitions is illustrated in Figure 10.3. Lemma 10.3 implies that the VRMMU can be solved by computing a minimum weight perfect matching in $\bar{G}$. As a result, Algorithm 4 solves the VRMMU and we get the following result.

Theorem 10.4. *The VRMMU can be solved in $\mathcal{O}(|L|^3)$ time.*

Algorithm 4: The VRMMU

Input: Operation sites $L^{\text{AM}} \subseteq L$, $L^{\text{PM}} \subseteq L$, vehicle depot d, and distances $\text{dist}\colon L \cup \{d\} \times L \cup \{d\} \to \mathbb{N}$

Output: Route partition $R \subseteq \mathcal{R}$ of minimal cost

1 construct $\bar{G} = (V_A \cup V_B, E)$ with $V_A \coloneqq \{a_\ell : \ell \in L^{\text{AM}}\}$, $V_B = V_B^1 \cup V_B^2$ where $V_B^1 \coloneqq \{b_\ell : \ell \in L^{\text{PM}}\}$ and $V_B^2 \coloneqq \{\emptyset_1, \ldots, \emptyset_{|L^{\text{AM}}|-|L^{\text{PM}}|}\}$, and $E = V_A \times V_B$

2 set $w(\{a_{\ell_1}, b_{\ell_2}\}) = c(\ell_1, \ell_2)$ and $w(\{a_{\ell_1}, \emptyset_i\}) = c(\ell_1, \emptyset)$ for all $a_{\ell_1} \in V_A$, $b_{\ell_2} \in V_B^1$, and $\emptyset_i \in V_B^2$

3 compute a minimum weight perfect matching $M \subseteq E$ in $\bar{G}$

4 set $R = \{(\ell_1, \emptyset) \in \mathcal{R} : \{a_{\ell_1}, \emptyset_i\} \in M\} \cup \{(\ell_1, \ell_2) \in \mathcal{R} : \{a_{\ell_1}, b_{\ell_2}\} \in M\}$

5 **return** R

Proof. The correctness of Algorithm 4 follows directly from Lemma 10.3. Concerning the running time of Algorithm 4, we can construct the graph $\bar{G}$ in $\mathcal{O}(|L|^2)$ time. Furthermore, we can compute a minimum weight perfect matching in $\bar{G}$ using the Hungarian method (Kuhn, 1955) which runs in $\mathcal{O}((|V_A| + |V_B|)^3) = \mathcal{O}(|L|^3)$ time. □

As a result of Theorem 10.4, we can solve the VRMMU in polynomial time. Moreover, as we are free to use any algorithm for the computation of a minimum weight perfect matching in $\bar{G}$, the results by Schwartz et al. (2005) yield faster randomized algorithms for the VRMMU. To conclude this section, we make an observation for the VRMMU with metric distances that can be used the reduce the size of problem instances.

Lemma 10.5. *Given a VRMMU instance defined by the MMU operation sites (L^{AM}, L^{PM}) with $L^{AM}, L^{PM} \subseteq L$, depot d, and metric distances $\text{dist}\colon L \cup \{d\} \times L \cup \{d\} \to \mathbb{N}$, there always exists an optimal route partition $R \subseteq \mathcal{R}$ with $(\ell, \ell) \in R$ for all $\ell \in L^{AM} \cap L^{PM}$.*

Proof. To prove the lemma's statement, let us assume the contrary, i.e., that there does not exist an optimal route partition $R \subseteq \mathcal{R}$ with $(\ell, \ell) \in R$ for all $\ell \in L^{\text{AM}} \cap L^{\text{PM}}$. Let $R^* \subseteq \mathcal{R}$ be an optimal route partition that contains a maximum number of routes of the form $(\ell, \ell) \in \mathcal{R}$ for $\ell \in L^{\text{AM}} \cap L^{\text{PM}}$. By assumption, there exists $\bar{\ell} \in L^{\text{AM}} \cap L^{\text{PM}}$ such that $(\bar{\ell}, \bar{\ell}) \notin R^*$. Thus, by the definition of a route partition, there must exist routes $(\ell_1, \bar{\ell}) \in R^*$ and $(\bar{\ell}, \ell_2) \in R^*$ that cover the morning and afternoon operations at $\bar{\ell}$. It can now be shown that a reassignment of these routes as shown in Figure 10.4 does not increase the cost of R^*, as

$$\begin{aligned}
c(\ell_1, \bar{\ell}) + c(\bar{\ell}, \ell_2) &= \text{dist}(d, \ell_1) + \text{dist}(\ell_1, \bar{\ell}) + \text{dist}(\bar{\ell}, d) + \text{dist}(d, \bar{\ell}) + \text{dist}(\bar{\ell}, \ell_2) + \text{dist}(\ell_2, d) \\
&= \text{dist}(d, \bar{\ell}) + \text{dist}(\bar{\ell}, d) + \text{dist}(d, \ell_1) + \text{dist}(\ell_1, \bar{\ell}) + \text{dist}(\bar{\ell}, \ell_2) + \text{dist}(\ell_2, d) \\
&\geq \text{dist}(d, \bar{\ell}) + \text{dist}(\bar{\ell}, d) + \text{dist}(d, \ell_1) + \text{dist}(\ell_1, \ell_2) + \text{dist}(\ell_2, d) \\
&= \text{dist}(d, \bar{\ell}) + \text{dist}(\bar{\ell}, \bar{\ell}) + \text{dist}(\bar{\ell}, d) + \text{dist}(d, \ell_1) + \text{dist}(\ell_1, \ell_2) + \text{dist}(\ell_2, d) \\
&= c(\bar{\ell}, \bar{\ell}) + c(\ell_1, \ell_2).
\end{aligned}$$

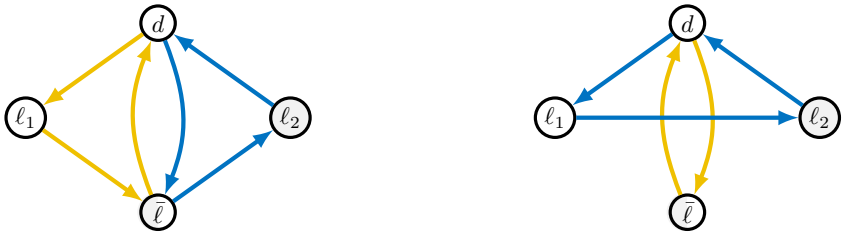


Fig. 10.4.: Reassignment of vehicle routes to maximize the number of routes that service the same site in the morning and afternoon session.

Therefore, we have that $\bar{R} = (R^* \setminus \{(\ell_1, \bar{\ell}), (\bar{\ell}, \ell_2)\}) \cup \{(\bar{\ell}, \bar{\ell}), (\ell_1, \ell_2)\}$ is a route partition with $c(\bar{R}) \leq c(R^*)$. Thus, $\bar{R}$ must be an optimal route partition with $|\{\ell \in L : (\ell, \ell) \in \bar{R}\}| > |\{\ell \in L : (\ell, \ell) \in R^*\}|$ which yields a contradiction and completes our proof. □

Applying Lemma 10.5, we can reduce the VRMMU with metric distances on sites $(L^{\text{AM}}, L^{\text{PM}})$ to the VRMMU with sites $(L^{\text{AM}} \setminus L^{\text{PM}}, L^{\text{PM}} \setminus L^{\text{AM}})$. If the distances are not metric or if we restrict the set of routes we may choose from, Lemma 10.5 does not hold in general.

In the next section, we extend the previous problem definition to multiple depots.

10.2 MMU Routing with Multiple Depots

In this section, we consider the vehicle routing problem for MMUs with multiple depots (mVRMMU) as an extension of the VRMMU which was limited to a single depot. Thus, let the set D denote the depots where the number of available MMUs at each depot $d \in D$ is given by $\nu_d \in \mathbb{N}$. Further, let $(L^{\text{AM}}, L^{\text{PM}})$ with $L^{\text{AM}} \subseteq L$, $L^{\text{PM}} \subseteq L$ be a pair of MMU operation sites which need to be serviced in the respective morning and afternoon session of the same day. Recall, that we assume w.l.o.g. that $|L^{\text{AM}}| \geq |L^{\text{PM}}|$. Moreover, we require that $\sum_{d \in D} \nu_d \geq |L^{\text{AM}}|$ to ensure the problem's feasibility. In contrast to Section 10.1, pairs of operation sites no longer uniquely define vehicle routes in the multi-depot setting. Instead, as we require vehicles to return to their starting depot at the end of each day, every tuple $(\ell_1, \ell_2, d) \in L^{\text{AM}} \times L^{\text{PM}} \times D$ uniquely defines the vehicle route $d \to \ell_1 \to \ell_2 \to d$ of length $c(\ell_1, \ell_2, d) := \text{dist}(d, \ell_1) + \text{dist}(\ell_1, \ell_2) + \text{dist}(\ell_2, d)$. Analogously to Section 10.1, we introduce the empty site $\emptyset$ to represent routes that service only the morning session, i.e., the tuple $(\ell, \emptyset, d)$ with $\ell \in L^{\text{AM}}$ and $d \in D$ encodes the route $d \to \ell \to d$ of length $c(\ell, \emptyset, d) := \text{dist}(d, \ell) + \text{dist}(\ell, d)$. We can now represent all vehicle routes by the set $\mathcal{R} := \{(\ell_1, \ell_2, d) : \ell_1 \in L^{\text{AM}} \cup \{\emptyset\}, \ell_2 \in L^{\text{PM}} \cup \{\emptyset\}, d \in D\}$. As a result, we need to adapt the definition of a route partition for multiple depots.

Definition 10.6. Let the pair of MMU operation sites $(L^{\text{AM}}, L^{\text{PM}})$ with $L^{\text{AM}} \subseteq L$, $L^{\text{PM}} \subseteq L$ be given. We call a set of the vehicle routes $R \subseteq \mathcal{R}$ with $|R| = |L^{\text{AM}}|$ a *route partition* if and only if for all $\ell \in L^{\text{AM}}$ it holds that $|\{(\ell, \ell_2, d) \in R\}| = 1$, for all $\ell \in L^{\text{PM}}$ it holds that

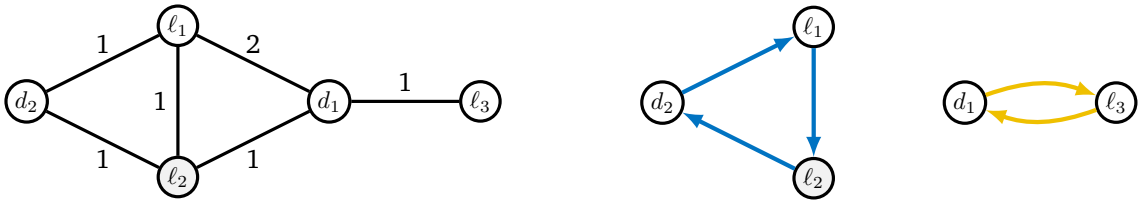


Fig. 10.5.: Example of a route partition: (a) mVRMMU instance, (b) route partition.

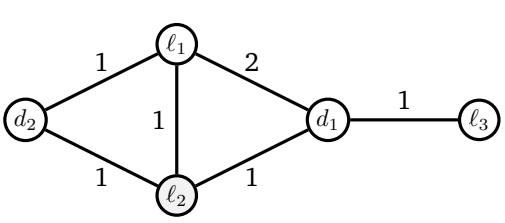
$|\{(\ell_1, \ell, d) \in R\}| = 1$, and for all $d \in D$ it holds that $|\{(\ell_1, \ell_2, d) \in R\}| \leq \nu_d$. The *cost* of a route partition is defined by the length of its routes, i.e., $c(R) = \sum_{(\ell_1,\ell_2,d)\in R} c(\ell_1, \ell_2, d)$.

An example of a route partition for multiple depots can be found in Figure 10.5. As in the one depot setting, each route partition must consist of $|L^{\text{PM}}|$ routes that service the morning and the afternoon session sites and $|L^{\text{AM}}| - |L^{\text{PM}}|$ routes that service only the morning session. We can now formally define the vehicle routing problem for MMUs with multiple depots.

Definition 10.7 (mVRMMU)**.** Given a pair of MMU operation sites $(L^{\text{AM}}, L^{\text{PM}})$ with $L^{\text{AM}} \subseteq L$ and $L^{\text{PM}} \subseteq L$, vehicle depots $d \in D$ with $\nu_d \in \mathbb{N}$ available vehicles, and distances $\text{dist}\colon L \cup D \times L \cup D \to \mathbb{N}$ between them, the *vehicle routing problem for MMUs with multiple depots* (mVRMMU) asks for a route partition of minimum cost.

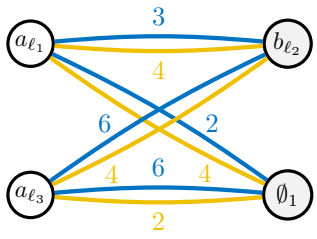
Similar to the previous section, we can reduce the mVRMMU to a weighted matching problem in a complete bipartite multi-graph. However, as the number of vehicles per depot is limited, we get additional budget constraints that limit the number of edges in a matching that correspond to the same depot. Specifically, we construct a bipartite graph $\bar{G} = (V_A \cup V_B, E)$ with nodes $V_A \coloneqq \{a_\ell : \ell \in L^{\text{AM}}\}$ and $V_B = V_B^1 \cup V_B^2$ where $V_B^1 \coloneqq \{b_\ell : \ell \in L^{\text{PM}}\}$ and $V_B^2 \coloneqq \{\emptyset_1, \ldots, \emptyset_{|L^{\text{AM}}|-|L^{\text{PM}}|}\}$. The set of edges is given by $E = \bigcup_{d\in D} E_d$, where $E_d = V_A \times V_B$ models the set of routes starting and ending in depot $d \in D$. As such, $\bar{G}$ is a multi-graph that contains each edge from the one depot setting $|D|$ times. Note, that the partition of E can be interpreted as an edge coloring with $|D|$ colors that assigns each edge $e \in E_d$ the color $d \in D$. For each $\{a_{\ell_1}, b_{\ell_2}\} \in E_d$, we choose the edge weight $w(\{a_{\ell_1}, b_{\ell_2}\}) = c(\ell_1, \ell_2, d)$ and set $w(\{a_{\ell_1}, \emptyset_i\}) = c(\ell_1, \emptyset, d)$ for $\{a_{\ell_1}, \emptyset_i\} \in E_d$. An illustration for the construction of the multi-graph $\bar{G}$ visualizing the edge partitioning as an edge coloring can be found in Figure 10.6. We can now show the following correspondence.

Lemma 10.8. *Every perfect matching $M \subseteq E$ in the multi-graph $\bar{G} = (V_A \cup V_B, E)$ with $E = \bigcup_{d\in D} E_d$ that satisfies $|M \cap E_d| \leq \nu_d$ for all $d \in D$ corresponds to a route partition $R \subseteq \mathcal{R}$ with $c(R) = w(M)$ and vice versa.*



$L^{\text{AM}} = \{\ell_1, \ell_3\},\ L^{\text{PM}} = \{\ell_2\},\ \nu_{d_1} = \nu_{d_2} = 1$

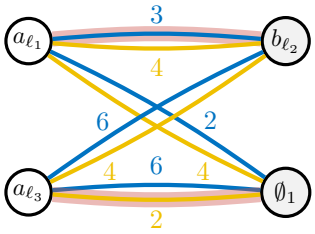
(a)



$\bar{G} = (V_A \cup V_B, E_{d_1} \cup E_{d_2})$

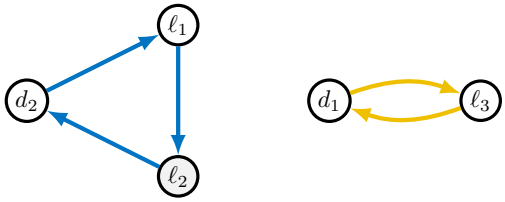
(b)

Fig. 10.6.: Exemplary construction of $\bar{G}$: (a) mVRMMU instance where unspecified distances correspond to the length of a shortest path, (b) corresponding multi-graph $\bar{G}$.



$M = \{\{a_{\ell_1}, b_{\ell_2}\}, \{a_{\ell_3}, \emptyset_1\}\},\ w(M) = 5$

(a)



$R = \{(\ell_1, \ell_2, d_2), (\ell_3, \emptyset, d_1)\},\ c(R) = 5$

(b)

Fig. 10.7.: Matchings and route partitions for mVRMMU instance in Figure 10.6 : (a) Perfect bipartite matching in $\bar{G}$, (b) corresponding route partition of identical cost.

Proof. Let $M \subseteq E$ be a perfect matching in $\bar{G} = (V_A \cup V_B, E)$ with $E = \bigcup_{d \in D} E_d$ that satisfies $|M \cap E_d| \leq \nu_d$ for all $d \in D$. We construct a route partition $R = R_1 \cup R_2$ with $R_1 \coloneqq \{(\ell_1, \emptyset, d) \in \mathcal{R} : \{a_{\ell_1}, \emptyset_i\} \in M \cap E_d\}$ and $R_2 \coloneqq \{(\ell_1, \ell_2, d) \in \mathcal{R} : \{a_{\ell_1}, b_{\ell_2}\} \in M \cap E_d\}$. As M is a perfect matching in G, it follows that $|R| = |M| = |L^{\text{AM}}|$. For each site $\ell \in L^{\text{AM}}$, it holds that $|\{(\ell, \ell_2, d) \in R\}| = |\delta_M(a_\ell)| = 1$. Moreover, we have that $|\{(\ell_1, \ell, d) \in R\}| = |\delta_M(b_\ell)| = 1$ for all $\ell \in L^{\text{PM}}$. For all depots $d \in D$ we furthermore have by assumption that $|\{(\ell_1, \ell_2, d) \in R\}| = |M \cap E_d| \leq \nu_d$ which proves that R is a route partition. Concerning the cost of R, we have

$$\begin{aligned} c(R) &= \sum_{d \in D} \sum_{(\ell_1, \ell_2, d) \in R} c(\ell_1, \ell_2, d) \\ &= \sum_{d \in D} \sum_{e \in M \cap E_d} w(e) = w(M). \end{aligned}$$

The converse direction can be shown analogously. □

An example illustrating how a perfect matching in $\bar{G}$ corresponds to a route partition can be found in Figure 10.7. As the size of $\bar{G}$ is polynomially bounded in the encoding length of an mVRMMU instance, Lemma 10.8 allows us to reduce the mVRMMU to a budgeted perfect matching problem in a bipartite edge colored multi-graph that we define as follows.

Definition 10.9. (BCBPM). Let $G = (V_A \cup V_B, E)$ be a bipartite multi-graph with $|V_A| = |V_B|$ and edge coloring $E = E_1 \cup \cdots \cup E_k$. Moreover, let $w : E \to \mathbb{N}$ be an edge weight function, and $B_i \in \mathbb{N}$ be the budget per color class E_i for $i \in \{1, \ldots, k\}$. The *budgeted colored bipartite perfect matching problem* (BCBPM) asks for a perfect matching $M \subseteq E$ of minimum weight $w(M) = \sum_{e \in M} w(e)$ such that the number of edges in each color class does not exceed the given budget, i.e., $|E_i \cap M| \leq B_i$ for all $i \in \{1, \ldots, k\}$.

The BCBPM generalizes the mVRMMU to general bipartite multi-graphs and independent edge weights as the edge weights in the mVRMMU are linked through the common trips from and to the depots. We show that the BCBPM is strongly $\mathcal{NP}$-hard via a reduction from (3,B2)-SAT that is inspired by Darmann et al. (2011).

Theorem 10.10. *The decision version of the BCBPM on planar graphs with uniform edge weights and budgets is strongly $\mathcal{NP}$-complete.*

Proof. First, let us recall that (3,B2)-SAT is the strongly $\mathcal{NP}$-complete symmetric specialization of the 3-SAT problem in which every clause consists of exactly 3 literals and every literal occurs exactly twice (Berman et al., 2003). This means that every variable occurs exactly four times – twice as positive and twice as negative literal. Let $\mathcal{I}$ be a (3,B2)-SAT instance with variables $X = \{x_1, \ldots, x_n\}$ and clauses $C = \{C_1, \ldots, C_m\}$. We construct an instance $\mathcal{I}'$ of the BCBPM as follows. The graph G is composed of n 4-cycles $e_1^i e_2^i e_3^i e_4^i$; one for every variable $x_i \in X$. Every perfect matching in G must contain either $T_i = \{e_1^i, e_3^i\}$ or $F_i = \{e_2^i, e_4^i\}$ for every variable $x_i \in X$ and we associate T_i with setting $x_i =$ True and F_i with setting $x_i =$ False; see Figure 10.8(a). We use unit edge weights and construct a partition of E using one color class E_j per clause $C_j \in C$. For every variable $x_i \in X$, let $C_{\bar{i}_1}, C_{\bar{i}_2}$ be the clauses containing literal $\bar{x}_i$ and C_{i_1}, C_{i_2} be the clauses containing literal x_i. We color the edges $e_1^i, e_3^i \in T_i$ in the colors $\bar{i}_1$ and $\bar{i}_2$, respectively. This way, choosing T_i (which corresponds to setting $x_i =$ True) counts towards the budget of every clause containing the literal $\bar{x}_i$. Analogously, we color the edges $e_2^i, e_4^i \in F_i$ in the colors i_1 and i_2, respectively. By construction, the number of edges $|M \cap E_j|$ indicates the number of unsatisfied literals in clause C_j for all $j \in \{1, \ldots, m\}$ and we thus set the budget $B_j = 2$ for all $j \in \{1, \ldots, m\}$. To clarify the construction of G, we refer to the example in Figure 10.8(b). We show that $\mathcal{I}$ is a Yes-instance if and only if $\mathcal{I}'$ has a perfect matching of weight $2n$.

Let $\mathcal{I}$ be a Yes-instance and x^* a satisfying truth assignment. We construct a perfect matching M^* with weight $w(M^*) = 2n$ as follows: For every $i \in \{1, \ldots, n\}$ pick T_i if $x_i^* =$ True and F_i otherwise. Obviously, the resulting matching M^* is perfect and has weight $2n$. Hence, it remains to show that the m budget constraints are satisfied. Assume the contrary, i.e., there exists a clause $C_j \in C$ such that $|E_j \cap M^*| = 3$. But this would imply that x^* did not satisfy clause C_j which yields a contradiction. The other direction can be shown analogously.

As we can check the weight and feasibility of a given budgeted colored perfect matching in $\mathcal{O}(|E|)$ time, the decision version of the BCBPM is obviously in $\mathcal{NP}$ and the problem's strong $\mathcal{NP}$-completeness follows. □

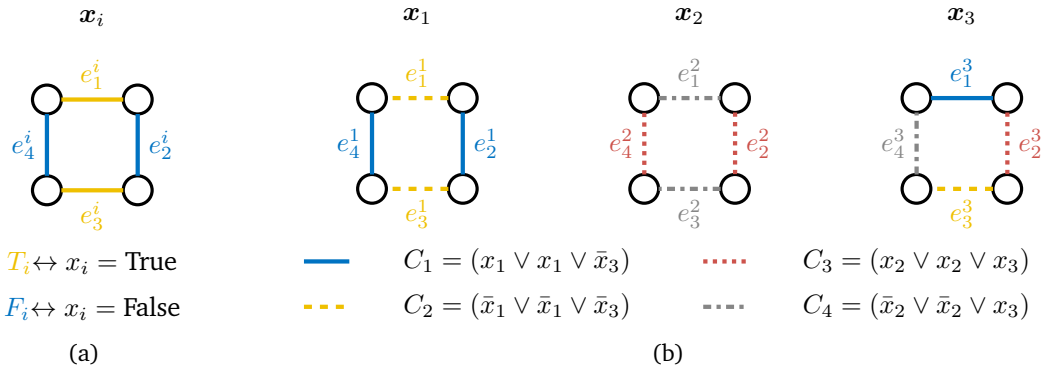


Fig. 10.8.: (a) Encoding of variables as 4-cycles, and (b) example of the construction of G for the reduction in Theorem 10.10.

The graph constructed in reduction above is not a complete bipartite multi-graph and therefore implicitly imposes restrictions on the set of routes we may choose from. Hence, we only get the following result.

Corollary 10.11. *The mVRMMU restricted to a subset of routes $\mathcal{R}' \subseteq \mathcal{R}$ is strongly $\mathcal{NP}$-hard, even if the number of MMUs per depot is uniform and the distances are metric.*

We can extend the construction in reduction for Theorem 10.10 to a complete bipartite multi-graph by adding all missing edges and setting their weights to 2 such that they cannot be in a perfect matching of weight $2n$. Unfortunately, this results in a structure of the edge weighs that cannot be achieved in the mVRMMU as we can infer from the following example. Consider the two-colored 4-cycle in Figure 10.9(a). To extend this 4-cycle to a complete bipartite multi-graph, we have to duplicate each edge as shown in Figure 10.9(b) (additional edges are dashed). The additional (dashed) edges must have a weight greater than the weight of the original edges to ensure that they cannot be chosen. In particular, each additional (dashed) edge must have a higher weight than their parallel counterpart. Keeping in mind that the edge weights must be defined as the route lengths in an mVRMMU instance of the form shown in Figure 10.9(c), we get the following requirements:

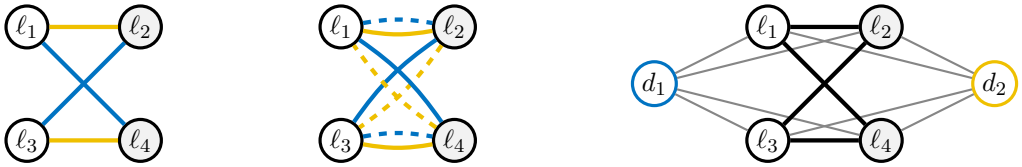
$$\text{dist}(d_2, \ell_1) + \text{dist}(\ell_2, d_2) < \text{dist}(d_1, \ell_1) + \text{dist}(\ell_2, d_1) \tag{10.1}$$

$$\text{dist}(d_1, \ell_1) + \text{dist}(\ell_4, d_1) < \text{dist}(d_2, \ell_1) + \text{dist}(\ell_4, d_2) \tag{10.2}$$

$$\text{dist}(d_2, \ell_3) + \text{dist}(\ell_4, d_2) < \text{dist}(d_1, \ell_3) + \text{dist}(\ell_4, d_1) \tag{10.3}$$

$$\text{dist}(d_1, \ell_3) + \text{dist}(\ell_2, d_1) < \text{dist}(d_2, \ell_3) + \text{dist}(\ell_2, d_2). \tag{10.4}$$

Adding all inequalities (10.1) – (10.4) yields a contraction, showing that the reduction from Theorem 10.10 does not extend the general mVRMMU. As a result, we do not obtain an $\mathcal{NP}$-hardness result for the general mVRMMU. Still, if we allow the distances between sites to depend on the depot a vehicle left from, we get the following.



(a) (b) (c)

Fig. 10.9.: (a) Example of 4-cycle used in the proof of Theorem 10.10. (b) Completion of 4-cycle by adding missing edges. (c) Corresponding mVRMMU instance with two depots.

Corollary 10.12. *The mVRMMU with depot-dependent distances* $\mathrm{dist}\colon L \cup D \times L \cup D \times D \to \mathbb{N}$ *is strongly* $\mathcal{NP}$*-hard, even if the number of MMUs per depot is uniform.*

We study a generalization of the BCBPM in Part III of this thesis and show that there are pseudo-polynomial dynamic programs for certain special cases. Whether or not the mVRMMU is $\mathcal{NP}$-hard remains open. To solve the mVRMMU, we resort to a straightforward mathematical programming formulation. Specifically, let $x_r \in \{0,1\}$ for $r \in \mathcal{R}$ be a binary variable indicating whether route r is part of the route partition $R \subseteq \mathcal{R}$. Then we can formulate the mVRMMU as follows

$$
\begin{aligned}
(\mathrm{mVR}) \quad \min_{x} \quad & \sum_{r \in \mathcal{R}} c(r)\, x_r && && \text{(10.5a)}\\
\text{s.t.} \quad & \sum_{r \in \mathcal{R}} x_r = |L^{\mathrm{AM}}| && && \text{(10.5b)}\\
& \sum_{(\ell,\ell_2,d) \in \mathcal{R}} x_r = 1 && \forall \ell \in L^{\mathrm{AM}} && \text{(10.5c)}\\
& \sum_{(\ell_1,\ell,d) \in \mathcal{R}} x_r = 1 && \forall \ell \in L^{\mathrm{PM}} && \text{(10.5d)}\\
& \sum_{(\ell_1,\ell_2,d) \in \mathcal{R}} x_r \le \nu_d && \forall d \in D && \text{(10.5e)}\\
& x_r \in \{0,1\} && \forall r \in \mathcal{R}. && \text{(10.5f)}
\end{aligned}
$$

In formulation (mVR), constraints (10.5b) – (10.5e) ensure that the selected set of vehicle routes $R = \{r \in \mathcal{R} : x_r = 1\}$ is indeed a route partition while the objective (10.5a) clearly minimizes the cost $c(R)$.

In the subsequent chapter, we evaluate the P3MMU in a case study. To generate the required input data, we use the agent-based simulation model SiM-Care from Part I.

11 Case Study: Optimized Operation of MMUs

In this case study, we showcase the applicability of our three-phased optimization approach P3MMU for the rural primary care system in the northern Eifel region of Germany that we previously modeled in SiM-Care; compare Chapter 5. To that end, Section 11.1 elaborates on the design of our set of test instances. Subsequently, we compute strategic MMU operation plans by solving the SMMU as well as the rSMMU in Section 11.2. In Section 11.3, we partition the obtained strategic MMU operation plans into tactical MMU operation plans by solving the TPMMU. To obtain the actual vehicle routes, we solve the VRMMU for the computed tactical MMU operation plans and a single depot in Section 11.4. Finally, we illustrate how the agent-based simulation tool SiM-Care from Part I can be used to evaluate optimized MMU operation plans in Section 11.5.

11.1 Test Instances

The primary care system that provides the template for our test instances comprises three predominantly rural municipalities in Western Germany. In the following, we successively consider the modeling of the practices P, potential MMU operation sites L, depot d, and patient demand origins V.

11.1.1 Practices

According to data provided by the local department of public health for the year 2017, there are 20 primary care physicians with health insurance accreditation in the considered primary care system. All physicians in the system operate in clinical sessions according to a weekly recurring schedule. The official consultation hours of each clinical session are publicly available from the Association of Statutory Health Insurance Physicians Nordrhein (Kassenärtzliche Vereinigung Nordrhein, 2019). In addition to the official consultation hours, we assume that the first hour after the end of each clinical session serves as a buffer during which physicians no longer accept new patients, but continue treating existing ones. To estimate each physician's weekly treatment capacity, we divide the total weekly consultation time (including buffers) by the average primary care physician's consultation time of $7.6\,\mathrm{min}$ as reported for Germany in Irving et al. (2017). After aggregating physicians that work in joint

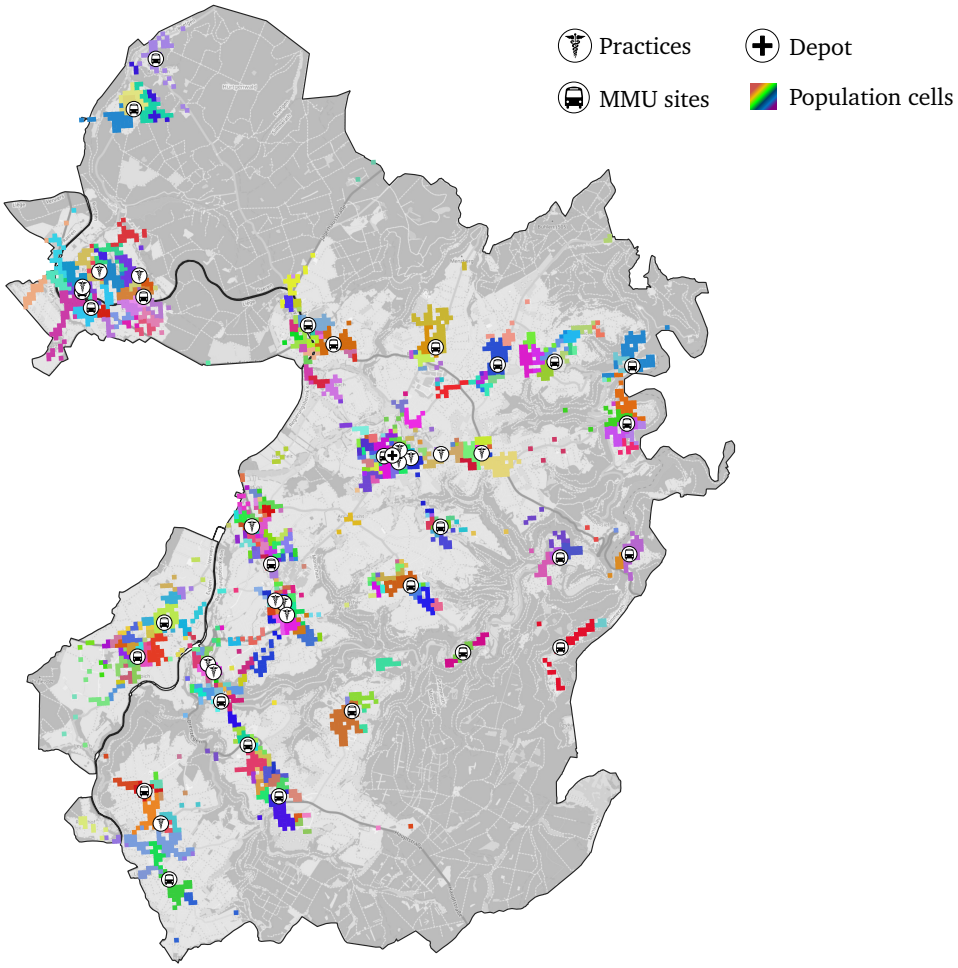


Fig. 11.1.: Locations of the practices, potential MMU locations, depot, and non-aggregated population cells clustered according to consideration sets for $\Delta = 6\,\mathrm{km}$.[6]

practices and ceiling the derived treatment capacities, this yields our set of $|P| = 16$ practices with treatment capacities $\bar{b}_p \in [206, 602]$ for all $p \in P$; see Figure 11.1.

11.1.2 MMU Operation Sites

Concerning the potential MMU operation sites L, we evenly distribute $|L| = 28$ sites among the agglomerations of the considered municipalities; compare Figure 11.1. Under the assumption that MMUs operate Monday to Friday in a morning and afternoon session, we set $b_\ell = 10$ for all sites $\ell \in L$. The duration of an MMU session is assumed to be $2.5\,\mathrm{h}$ which is slightly below the average session duration (without buffers) of $3.36\,\mathrm{h}$ observed for the physicians in the considered primary care system. Just as for practices, we do anticipate a buffer of one hour after each MMU sessions. By dividing the duration of an MMU session (including buffer) by the average German primary care physician's consultation time of

[6] Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

$7.6\,\mathrm{min}$ (Irving et al., 2017), we end up with a treatment capacity of $\hat{b} = 28$ per operated MMU session. With regard to the operation cost of MMUs, we assume that all sites are equally expensive to set up and that opening a new site is twice as undesirable as operating a weekly MMU session. Thus, we choose the setup cost $c_\ell = 2$ for all sites $\ell \in L$ and set the cost per operated MMU session to $\hat{c} = 1$.

11.1.3 Depot

With regard to the stationing of the MMUs, we assume that all vehicles start and end their routes in the region's only hospital; compare Figure 11.1.

11.1.4 Patient Demand Origins

To model the patient demand origins V, we rely on the population data determined by the latest German census conducted in 2011 (Information und Technik Nordrhein-Westfalen, 2016). The census reports a total population of $35{,}542$ for the considered primary care system, specified at a resolution of $2{,}754$ population cells measuring one hectare each. To determine the consideration set of each population cell, we construct a street graph for the considered region based on map data from OpenStreetMap (OpenStreetMap contributors, 2019) using OSMnx (Boeing, 2017). The centers of the population cells and the treatment facilities are mapped to their respective closest node in the street network (as the crow flies), and we compute the driving distances between all population cells and treatment facilities along the street network. The consideration set of each population cell is then defined as all treatment facilities that can be reached within a maximum driving distance $\Delta \in \mathbb{N}$ in kilometers. Thereby, we order consideration sets according to the distance between a treatment facility and the center of the population cell. As a last step, we aggregate all population cells with identical (including order) consideration sets to obtain the set of demand origins V. We note, that this aggregation and thus also the resulting set of patient demand origins V depends on the choice of the parameter $\Delta \in \mathbb{N}$. The non-aggregated population cells clustered according to their consideration sets for $\Delta = 6\,\mathrm{km}$ are exemplary shown in Figure 11.1.

Next, we consider the steerable and unsteerable patient demands at each demand origin. As empirical data concerning the primary care demand at each demand origin is unavailable, we have to rely on simulation to obtain rough estimates. Specifically, we use the baseline scenario of the considered primary care system in the hybrid agent-based simulation tool SiM-Care from Part I that we introduced in Section 5.1 to obtain the number of primary care visits $q_v^i \in \mathbb{N}$ per demand origin $v \in V$ for every week $i \in \{1, \dots 52\}$ in a one year time horizon. Unsteerable patient demands, as considered here, are not representable in SiM-Care and we therefore assume that a fixed percentage $\omega \in [0, 1]$ of the simulated primary care visits can be attributed to unsteerable patient demands.

In the deterministic setting of the SMMU, we then choose the patient demands at each demand origin $v \in V$ proportionately to the rounded average simulated demand $\bar{q}_v \coloneqq \text{round}\left(\frac{1}{52}\sum_{i=1}^{52} q_v^i\right)$ where $\text{round}\colon \mathbb{R} \to \mathbb{Z}$ denotes the rounding function $\text{round}(x) \coloneqq \lfloor x + 0.5 \rfloor$. Specifically, we set

$$u_v = \text{round}(\omega\, \bar{q}_v) \quad \text{and} \quad d_v = \bar{q}_v - u_v.$$

In the uncertain setting of the rSMMU, we choose the lower and upper bounds for the patient demands at each demand origin proportionately to the minimum and maximum simulated demands. Formally, this translates into setting

$$\sigma_v = \text{round}\left(\omega \min_{1 \le i \le 52}\{q_v^i\}\right) \quad \text{and} \quad \alpha_v = \min_{1 \le i \le 52}\{q_v^i\} - \sigma_v$$

for the lower bound of the unsteerable and steerable demands σ_v and α_v at $v \in V$, and

$$\tau_v = \text{round}\left(\omega \max_{1 \le i \le 52}\{q_v^i\}\right) \quad \text{and} \quad \beta_v = \max_{1 \le i \le 52}\{q_v^i\} - \tau_v$$

for the respective upper bounds τ_v and β_v at $v \in V$.

The budget parameters for the budgeted uncertainty sets are determined by the maximum simulated total demand among all weeks, i.e., we choose

$$\Gamma_2 = \text{round}\left(\omega \max_{1 \le i \le 52} \sum_{v \in V} q_v^i\right) \quad \text{and} \quad \Gamma_1 = \left(\max_{1 \le i \le 52} \sum_{v \in V} q_v^i\right) - \Gamma_2.$$

For our test instances, we consider the patient demand origins obtained by choosing a maximum driving distance $\Delta \in \{6, 7, \ldots, 11\}$ and by varying the share of the unsteerable patient demands $\omega \in \{0.2, 0.25, \ldots, 0.45\}$. Table 11.1 summarizes the characteristics of the resulting 36 test instances. We remark, that the total average demand $\sum_{v \in V} d_v + u_v$ as well as the total worst case demand $\sum_{v \in V} \beta_v + \tau_v$ in each instance depends on the choice of $\Delta \in \mathbb{N}$ as a result of the aggregation of population cells. For our set of test instances, the total average demand is relatively robust towards changes in Δ, while the total worst case demand increases with Δ due to the resulting larger number of demand origins $|V|$.

11.2 Study Phase 1

The computational study of Phase 1 focuses on the SMMU as well as the rSMMU with budgeted and interval uncertainty sets introduced in Sections 8.1 and 8.2, respectively. Based on the previously introduced set of realistic test instances, we compare the cost and quality of the strategic MMU operation plans resulting from the three different approaches and investigate the so-called price of robustness. To that end, we first describe the study

Tab. 11.1.: Test instances with their characteristics.

Instance	Δ	ω	$\lvert V\rvert$	$\lvert L\rvert$	$\lvert P\rvert$	$\sum d+u$	$\sum \beta+\tau$	Γ_1	Γ_2
1	6	0.2	417	28	16	3,888	6,400	3,233	808
2	6	0.25	417	28	16	3,888	6,400	3,031	1,010
3	6	0.3	417	28	16	3,888	6,400	2,829	1,212
4	6	0.35	417	28	16	3,888	6,400	2,627	1,414
5	6	0.4	417	28	16	3,888	6,400	2,425	1,616
6	6	0.45	417	28	16	3,888	6,400	2,223	1,818
7	7	0.2	462	28	16	3,886	6,517	3,233	808
8	7	0.25	462	28	16	3,886	6,517	3,031	1,010
9	7	0.3	462	28	16	3,886	6,517	2,829	1,212
10	7	0.35	462	28	16	3,886	6,517	2,627	1,414
11	7	0.4	462	28	16	3,886	6,517	2,425	1,616
12	7	0.45	462	28	16	3,886	6,517	2,223	1,818
13	8	0.2	506	28	16	3,885	6,668	3,233	808
14	8	0.25	506	28	16	3,885	6,668	3,031	1,010
15	8	0.3	506	28	16	3,885	6,668	2,829	1,212
16	8	0.35	506	28	16	3,885	6,668	2,627	1,414
17	8	0.4	506	28	16	3,885	6,668	2,425	1,616
18	8	0.45	506	28	16	3,885	6,668	2,223	1,818
19	9	0.2	546	28	16	3,888	6,829	3,233	808
20	9	0.25	546	28	16	3,888	6,829	3,031	1,010
21	9	0.3	546	28	16	3,888	6,829	2,829	1,212
22	9	0.35	546	28	16	3,888	6,829	2,627	1,414
23	9	0.4	546	28	16	3,888	6,829	2,425	1,616
24	9	0.45	546	28	16	3,888	6,829	2,223	1,818
25	10	0.2	576	28	16	3,887	6,938	3,233	808
26	10	0.25	576	28	16	3,887	6,938	3,031	1,010
27	10	0.3	576	28	16	3,887	6,938	2,829	1,212
28	10	0.35	576	28	16	3,887	6,938	2,627	1,414
29	10	0.4	576	28	16	3,887	6,938	2,425	1,616
30	10	0.45	576	28	16	3,887	6,938	2,223	1,818
31	11	0.2	583	28	16	3,889	6,960	3,233	808
32	11	0.25	583	28	16	3,889	6,960	3,031	1,010
33	11	0.3	583	28	16	3,889	6,960	2,829	1,212
34	11	0.35	583	28	16	3,889	6,960	2,627	1,414
35	11	0.4	583	28	16	3,889	6,960	2,425	1,616
36	11	0.45	583	28	16	3,889	6,960	2,223	1,818

design in Section 11.2.1 before the actual computational results of the study are presented in Section 11.2.2.

11.2.1 Implementation and Computational Setup

In our computational study, we implemented all mathematical programs in Java using OpenJDK 11 (Oracle, 2018) and the CPLEX 12.8 Java API (IBM, 2018). The CPLEX optimizer is restricted to one thread and all other CPLEX parameters are left at their default settings. Instances of the SMMU are solved using formulation (Det-B) and instances of the rSMMU with interval uncertainty sets are solved using formulation (RobI-B). To solve instances of the rSMMU with budgeted uncertainty sets, we use formulation (RobΓ-B). The separation

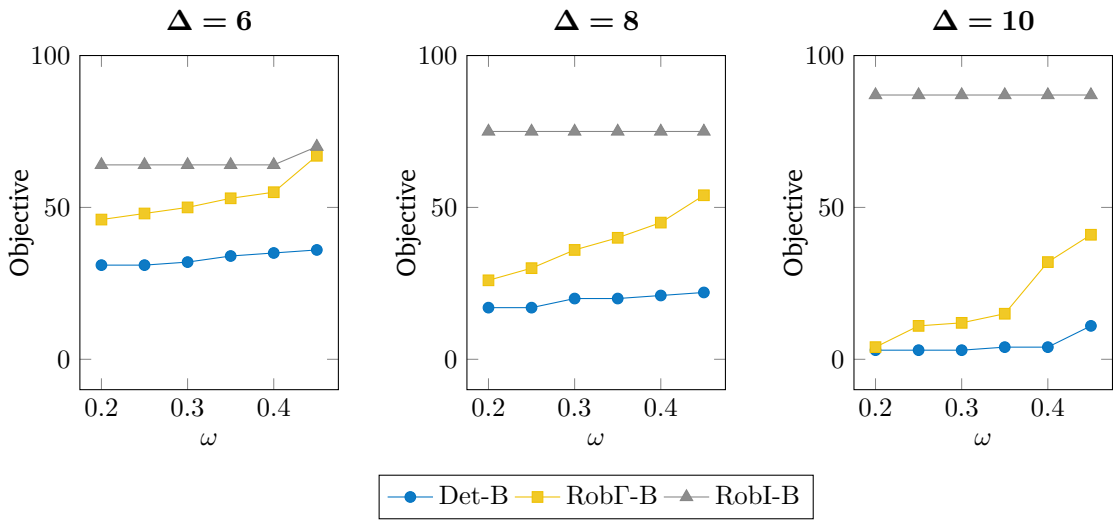


Fig. 11.2.: Objective function values for fixed $\Delta \in \{6, 8, 10\}$ and varying $\omega \in \{0.2, \ldots, 0.45\}$.

problems in (Det-B) and (RobI-B) are solved using the LP-formulation from Appendix A.2, and we integrate the separation procedure directly into the branch-and-bound scheme using lazy constraint callbacks. The separation problem in (RobΓ-B) is solved using formulation (Sep) and we call the separation procedure only after the restricted master problem is solved to optimality. Note, that the separation for (RobΓ-B) cannot be integrated into the branch-and-bound scheme as lazy constraint callbacks do not allow for the introduction of new variables. As we cannot ensure that Assumption 1 holds for our test instances, we apply the explicit enforcement from Appendix A.1. All computational experiments were performed on a cluster of machines running Ubuntu 18.04 with an Intel(R) Core(TM) i9-9900 CPU @ 3.10 GHz and 32 GB DDR4-Non-ECC main memory. We restrict each individual job to one physical core and 3.5 GB main memory. All instances were solved to optimality (CPLEX default MIP gap tolerance 10^{-4}) and all running times are reported in CPU seconds.

11.2.2 Computational Results

The optimal objective values and CPU times of (Det-B), (RobΓ-B), and (RobI-B) for the 36 test instances are summarized in Table 11.2. In the following, we discuss these results with a focus on the impact of the percentage of unsteerable patient demands $\omega \in [0, 1]$ and the patients' maximum driving distance $\Delta \in \mathbb{N}$. Furthermore, we investigate the *price of robustness* – a term introduced by Bertsimas and Sim (2004) that describes the additional cost of a robust solution compared to a non-robust solution that has to be payed for the protection against data uncertainties. To that end, we compare the objective values of the robust solutions for (RobΓ-B) and (RobI-B) to the objective values of the nominal solutions for (Det-B).

Examining the impact of the percentage of unsteerable patient demands, we visualize the objective function values of (Det-B), (RobΓ-B), and (RobI-B) for exemplary fixed $\Delta \in \mathbb{N}$ and

Tab. 11.2.: Computational results for Phase 1.

Instance			Objective			CPU [s]		
#	Δ	ω	Det-B	RobΓ-B	RobI-B	Det-B	RobΓ-B	RobI-B
1	6	0.2	31	46	64	6	32	5
2	6	0.25	31	48	64	8	44	3
3	6	0.3	32	50	64	7	45	5
4	6	0.35	34	53	64	5	96	5
5	6	0.4	35	55	64	3	139	8
6	6	0.45	36	67	70	5	93	12
7	7	0.2	25	32	68	10	26	3
8	7	0.25	26	37	68	6	38	5
9	7	0.3	27	39	68	5	26	6
10	7	0.35	29	43	68	7	26	7
11	7	0.4	30	48	68	5	57	3
12	7	0.45	31	52	68	3	18	4
13	8	0.2	17	26	75	8	73	12
14	8	0.25	17	30	75	6	42	4
15	8	0.3	20	36	75	9	51	9
16	8	0.35	20	40	75	9	41	7
17	8	0.4	21	45	75	7	44	34
18	8	0.45	22	54	75	4	19	6
19	9	0.2	9	13	81	2	18	1,678
20	9	0.25	10	15	81	4	16	683
21	9	0.3	11	21	81	5	29	1,936
22	9	0.35	11	27	81	4	30	1,847
23	9	0.4	12	32	81	3	49	1,069
24	9	0.45	13	47	81	4	29	24
25	10	0.2	3	4	87	4	14	4
26	10	0.25	3	11	87	2	10	6,944
27	10	0.3	3	12	87	4	10	5,993
28	10	0.35	4	15	87	2	8	8,577
29	10	0.4	4	32	87	2	16	1,126
30	10	0.45	11	41	87	4	10	191
31	11	0.2	0	0	88	2	8	10,026
32	11	0.25	0	7	88	2	9	4
33	11	0.3	0	8	88	2	12	12,815
34	11	0.35	0	15	88	2	20	12,925
35	11	0.4	0	32	88	2	12	4,968
36	11	0.45	7	41	88	4	13	1,403

varying $\omega \in [0, 1]$ in Figure 11.2. From the way we modeled the unsteerable patient demands, it is our expectation that a higher percentage of unsteerable patient demands leads to a higher objective value as a result of the associated loss of control over the patient demands. Looking at Figure 11.2, we can confirm this expectation regardless of the fixed maximum driving distance $\Delta \in \mathbb{N}$ and the considered setting. However, we can observe large differences in the degree of this effect. In the robust setting with interval uncertainty sets (RobI-B), the objective values are mostly unaffected by the choice of ω. This can be attributed to the conservatism of this approach, which leads to an overloading of the existing primary care system where the sheer level of demand seems to dominate the cost of the MMU operation plan. In the deterministic setting (Det-B), the influence of the percentage of the unsteerable

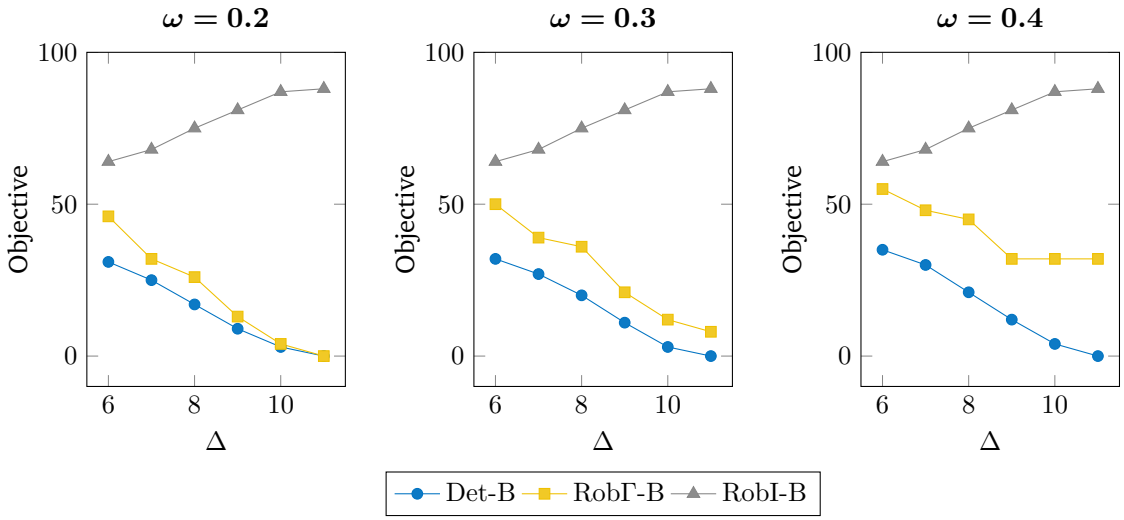


Fig. 11.3.: Objective function values for fixed $\omega \in \{0.2, 0.3, 0.4\}$ and varying $\Delta \in \{6, \ldots, 11\}$.

patient demands is more pronounced, yet still relatively weak. One explanation for this behavior is that the local level of unsteerable patient demands in this setting remains at a degree which can be mostly compensated by an appropriate reassignment of the steerable patient demands. The greatest impact of the percentage of unsteerable patient demands $\omega \in [0, 1]$ on the cost of the MMU operation plan can be observed in the robust setting with budgeted uncertainty sets for (RobΓ-B). This showcases, that the budgeted uncertainty sets succeed at limiting the total patient demand as opposed to interval uncertainty sets, while still accounting for local worst-cases as opposed to the deterministic setting. Concerning the price of robustness, we can confirm that the budgeted uncertainty sets manage to substantially lower the price of robustness compared to the interval uncertainty sets. Furthermore, we can observe that the price of robustness is lowest for small $\omega \in [0, 1]$ and increases with the percentage of the unsteerable patient demands. The entire set of figures for this evaluation can be found in Appendix A.3.

To analyze the impact of the patients' maximum driving distance, we visualize the objective function values of (Det-B), (RobΓ-B), and (RobI-B) for exemplary fixed $\omega \in [0, 1]$ and varying $\Delta \in \mathbb{N}$ in Figure 11.3. Intuitively, one would assume that a higher patients' maximum driving distance leads to a lower objective function value as a result of larger consideration sets which yield more flexibility in the assignment of steerable patient demands. However, looking at Figure 11.3 this assumption can only be verified for (Det-B) and (RobΓ-B). For the choice of interval uncertainty sets in (RobI-B), we can actually observe the opposite behavior. Although this might seem counterintuitive at first glance, we can explain this behavior by the aggregation of the population cells during the instance generation process as described in Section 11.1: An increase in Δ leads to more diverse consideration sets which, in turn, result in a higher number of demand origins $|V|$ as well as a higher total worst case patient demand $\sum_{v \in V} \beta_v + \tau_v$; compare Table 11.1. Paired with the previous observation that the objective value of (RobI-B) seems to be dominated by this worst case demand, an increase in the objective value is actually to be expected. In (RobΓ-B), we limit the total demand in the

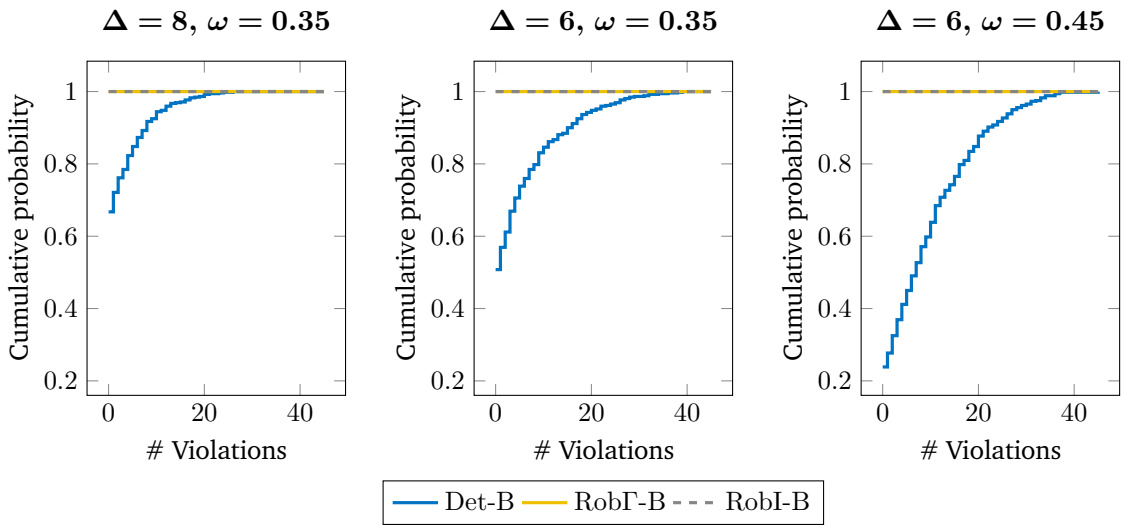


Fig. 11.4.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $(\Delta, \omega) \in \{(8, 0.35), (6, 0.35), (6, 0.45)\}$.

system through the use of budgeted uncertainty sets and thus can observe that the cost of MMU operation plans are decreasing in the patients' maximum driving distance. The extent of the savings associated with an increase in Δ, decreases as we increase the percentage of the unsteerable patient demands ω for both (Det-B) and (RobΓ-B). This makes perfect sense as an increase in ω necessarily results in a smaller percentage of steerable patient demands for which we can actually profit from the enlarged consideration sets. Concerning the price of robustness, also this collation of our results validates that the use of budgeted uncertainty sets reduces the price of robustness compared to the use of interval uncertainty sets. Moreover, the difference in the optimal solution values between (RobΓ-B) and (RobI-B) increases with the patients' maximum driving distance $\Delta \in \mathbb{N}$, which is partly due to the undesired increase in the total worst case demand resulting from the aggregation of population cells during instance generation. The entire set of figures for this evaluation can be found in Appendix A.3.

To justify why the price of robustness should be paid, we analyze the quality of the computed MMU operation plans. For this purpose, we reuse the SiM-Care model of the considered primary care system which we referred to in Section 11.1 to generate another set of patient demands for every week in a 10 year time horizon. Thereby, we do not aggregate the population cells specified by the German census such that we end up with 520 weekly demands for each of the 2,754 cells. To determine the unsteerable patient demands for each realization, we flip a biased coin where we set the success probability to the percentage of unsteerable patient demands $\omega \in [0, 1]$. We analyze for each MMU operation plan and each realization the minimum total number of violations of the treatment capacities. To clarify this metric, consider the following example: If 220 patients are assigned to a practice $p \in P$ with weekly treatment capacity $\bar{b}_p = 200$, this yields 20 violations.

Figure 11.4 shows the minimum total number of violations obtained by (Det-B), (RobΓ-B), and (RobI-B) for each of the 520 realizations and exemplary parameter choices. The first

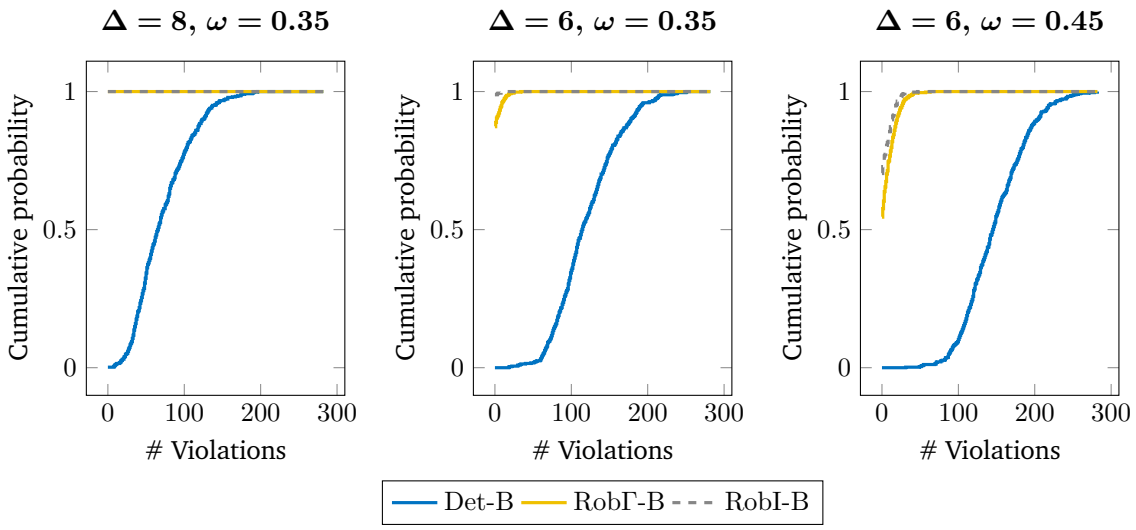


Fig. 11.5.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks and $(\Delta, \omega) \in \{(8, 0.35), (6, 0.35), (6, 0.45)\}$.

thing we want to emphasize, is that the solutions obtained using the robust models (RobΓ-B) and (RobI-B) are actually feasible for all 520 realizations as there are no violations. This feasibility of the robust solutions is not unique to the depicted parameter choices, but actually holds for all robust solutions we computed. Considering the MMU operation plans obtained using (Det-B) and deterministic average demands, we can observe violations for quite a few realizations. The number and extent of these violations depends on the percentage of unsteerable patient demands $\omega \in [0, 1]$ and the patients' maximum driving distance $\Delta \in \mathbb{N}$. Specifically, the quality of the deterministic solutions deteriorates as we decrease Δ and increase ω which seems reasonable as these are exactly the settings for which we observed the highest price of robustness. Nevertheless, we must note that the highest number of violations observed over all parameter settings and realizations is 45. Setting 45 violations in relation to the total mean demand of roughly 3,900, only 1 % of the demands cannot be accounted for by the deterministic solutions. The entire set of figures for this evaluation can be found in Appendix A.4.

Although each violation can potentially lead to an avoidable emergency room visit, this is an admittedly good performance of the deterministic solutions. A possible reason for this is the fact that SiM-Care does not feature an infectious model; compare Part I. Thus, the generated realizations do not show the local surges in demand resulting from the outbreak of an infectious disease. To include these local demand spikes into our evaluation, we mimic infectious outbreaks in a very simplistic manner: We select successively at random 5 of the 2,754 population cells as outbreak centers and double the demands of each population cell within a 1 km radius of the outbreak center (as the crow flies). Figure 11.5 shows the minimum total number of violations obtained by (Det-B), (RobΓ-B), and (RobI-B) for each of the 520 realizations under the presence of infections outbreaks and exemplary parameter choices. Looking at the results, we can observe that the local surges in demand lead to more violations in all solutions. Evidently, the deterministic solutions perform worst, while the two

robust approaches produce solutions of comparable quality. The entire set of figures for this evaluation can be found in Appendix A.5.

While running times are clearly not a focus of this study, we note that all instances of $(\mathrm{Det\text{-}B})$ were solved within 10 CPU seconds which is sufficiently quick for a planning problem at a strategic level and leaves room to consider even larger primary care systems. The instances of $(\mathrm{Rob\Gamma\text{-}B})$, which are probably the most interesting ones from an application point of view, are noticeably more challenging than the instances of $(\mathrm{Det\text{-}B})$. However, even the hardest instance solved within 139 CPU seconds which is more than reasonable for this kind of strategic application. Formulation $(\mathrm{RobI\text{-}B})$ is undoubtedly the most challenging among all considered formulations. Especially for higher values of $\Delta \in \mathbb{N}$, CPLEX struggles to close the MIP gap which results in a tailing-off phenomenon and running times of up to $12{,}925$ CPU seconds. Although this is considerably longer than the running times of the other formulations, even running times of this magnitude can still be deemed acceptable for a strategic planning problem. Moreover, we note that the running times for the rSMMU with interval uncertainty sets and high values of $\Delta \in \mathbb{N}$ can be substantially improved by using a variation of formulation (Det) instead of $(\mathrm{RobI\text{-}B})$ that we decided to omit as we consider this setting for reference rather than as a serious alternative for real-world application.

11.3 Study Phase 2

The computational study of Phase 2 investigates the solvability of the TPMMU for the strategic MMU operation plans computed in the previous section. Section 11.3.1 describes the design of this study and provides details on our implementation. The computational results are presented in Section 11.3.2.

11.3.1 Implementation and Computational Setup

We implemented all mathematical programs in Java using OpenJDK 11 (Oracle, 2018) and the CPLEX 12.8 Java API (IBM, 2018). The CPLEX optimizer is restricted to one thread and all other CPLEX parameters are left at their default settings. All Instances of the TPMMU are solved using the compact formulation (TP). As all practices in the considered primary care system are closed on Wednesday afternoons, we minimize the sum of the covering radii as suggested in Section 9.1. For each instance, we consider the set of the $2{,}754$ unaggregated population cells and compute distances as the crow flies. All computational experiments were performed on a cluster of machines running Ubuntu 18.04 with an Intel(R) Core(TM) i9-9900 CPU @ $3.10\,\mathrm{GHz}$ and $32\,\mathrm{GB}$ DDR4-Non-ECC main memory. We restrict each individual job to one physical core and $3.5\,\mathrm{GB}$ main memory. All instances were solved to optimality (CPLEX default MIP gap tolerance 10^{-4}) and all running times are reported in CPU seconds.

Tab. 11.3.: Computational results for Phase 2.

Inst.	Det-B				RobΓ-B				RobI-B			
#	$\lvert L_{m^S}\rvert$	$\lvert m^S\rvert$	Obj.	CPU	$\lvert L_{m^S}\rvert$	$\lvert m^S\rvert$	Obj.	CPU	$\lvert L_{m^S}\rvert$	$\lvert m^S\rvert$	Obj.	CPU
1	6	19	66.24	122	6	34	62.02	496	6	52	49.99	2
2	6	19	66.24	121	7	34	60.40	348	6	52	49.63	1
3	6	20	66.04	131	6	38	62.54	297	6	52	51.18	1
4	6	22	64.35	139	7	39	59.27	1,236	6	52	52.53	1
5	6	23	64.15	175	7	41	55.93	13	6	52	52.73	1
6	6	24	64.31	435	9	49	55.95	770	9	52	51.33	222
7	5	15	66.32	34	5	22	69.47	38	6	56	50.56	1
8	5	16	66.24	70	5	27	63.77	82	6	56	54.17	1
9	5	17	66.04	37	5	29	64.05	98	6	56	53.15	1
10	5	19	65.84	77	5	33	60.75	89	6	56	51.56	2
11	5	20	70.29	103	6	36	59.54	119	6	56	49.03	1
12	5	21	64.26	132	6	40	61.78	27	6	56	50.86	2
13	4	9	79.02	13	4	18	70.46	11	7	60	62.10	2
14	4	9	76.12	13	4	22	64.85	3	7	61	62.36	2
15	4	12	71.58	13	5	26	64.86	3	7	61	62.10	1
16	4	12	71.58	15	5	30	64.63	64	7	61	63.03	1
17	4	13	71.38	14	5	35	61.03	14	7	61	62.10	3
18	4	14	71.38	14	6	42	64.64	5	7	61	63.60	2
19	2	5	80.58	1	2	9	79.82	1	7	67	71.23	1
20	2	6	80.22	1	2	11	77.41	3	7	67	64.89	2
21	2	7	80.02	1	3	15	71.74	55	7	67	65.97	2
22	2	7	80.02	1	4	19	68.15	21	7	67	64.52	1
23	2	8	79.82	1	4	24	66.03	26	7	67	64.89	1
24	2	9	79.82	1	5	37	65.59	17	7	67	63.47	2
25	1	1	86.54	0	1	2	82.56	0	8	71	64.89	1
26	1	1	86.54	0	2	7	78.92	5	8	71	63.60	2
27	1	1	86.54	0	2	8	78.72	4	8	71	64.11	2
28	1	2	82.56	1	2	11	77.41	1	8	71	65.09	2
29	1	2	82.56	1	4	24	66.88	3	8	71	63.43	2
30	2	7	78.92	2	5	31	64.64	14	8	71	65.22	2
31	0	0	–	–	0	0	–	–	8	72	65.89	1
32	0	0	–	–	1	5	82.01	1	8	72	63.60	2
33	0	0	–	–	1	6	81.81	1	8	72	66.29	1
34	0	0	–	–	2	11	77.41	1	8	72	66.36	1
35	0	0	–	–	4	24	65.11	13	8	72	59.08	1
36	1	5	82.01	0	5	31	65.35	11	8	72	65.20	2

11.3.2 Computational Results

The optimal objective values and CPU times of (TP) for the 108 strategic MMU operation plans based on the 36 test instances are summarized in Table 11.3. Each row corresponds to the three strategic MMU operation plans that were obtained for the same test instance using formulations (Det-B), (RobΓ-B), and (RobI-B). For each strategic MMU operation plan m^S, we specify the number of set up operation sites $|L_{m^S}|$ where $L_{m^S} := \{\ell \in L : m^S_\ell > 0\}$ as well as the total number of scheduled sessions $|m^S| := \sum_{\ell \in L} m^S_\ell$. Note, that some strategic MMU operation plans from Phase 1 do not schedule any MMU sessions. For these plans, there is no

Phase 2 to be solved which we indicate by not providing an objective function value and CPU time in Table 11.3.

Comparing the objective function values for the 108 strategic MMU operation plans, we can observe that overall a higher number of set up sites and MMU sessions tends to lead to smaller covering radii which seems reasonable. However there are some exceptions from this relationship which does not come as a surprise, since the covering radius is not explicitly considered in optimization models from Phase 1. To overcome this drawback, we formalized the combined strategic and tactical planning problem for MMUs in Section 9.2 which is however not considered in this study.

Concerning running times, we notice that the strategic operations plans obtained by $(\mathrm{Rob}\Gamma\text{-B})$ seem to be more difficult to partition than the ones obtained through formulations $(\mathrm{Det}\text{-B})$ and $(\mathrm{Rob}\text{I-B})$. Nevertheless, all instances of the TPMMU were solved within $1{,}236$ CPU seconds which we deem acceptable for a planning problem at the tactical level.

11.4 Study Phase 3

To investigate Phase 3 within P3MMU, we compute minimum cost route partitions for the tactical MMU operations plans that were obtained from Phase 2. We detail our study design and implementation specifics in Section 11.4.1 before we present the computational results in Section 11.4.2.

11.4.1 Implementation and Computational Setup

For this study, we implemented all mathematical programs in Java using OpenJDK 11 (Oracle, 2018) and the CPLEX 12.8 Java API (IBM, 2018). The CPLEX optimizer is restricted to one thread and all other CPLEX parameters are left at their default settings. The VRMMU for each day in every tactical MMU operation plan are solved sequentially using Algorithm 4. The lengths of the vehicle routes are computed as the crow flies and we use a linear programming formulation to solve the minimum weight perfect matching problem in the constructed graphs. All computational experiments were performed on a cluster of machines running Ubuntu 18.04 with an Intel(R) Core(TM) i9-9900 CPU @ $3.10\,\mathrm{GHz}$ and $32\,\mathrm{GB}$ DDR4-Non-ECC main memory. We restrict each individual job to one physical core and $3.5\,\mathrm{GB}$ main memory. All instances were solved to optimality (CPLEX default MIP gap tolerance 10^{-4}) and all running times are reported in CPU seconds.

Tab. 11.4.: Computational results for Phase 3.

Inst.	Det-B				RobΓ-B				RobI-B			
#	$\nu(m^{\mathrm{T}})$	$\lvert m^{\mathrm{T}}\rvert$	Obj.	CPU	$\nu(m^{\mathrm{T}})$	$\lvert m^{\mathrm{T}}\rvert$	Obj.	CPU	$\nu(m^{\mathrm{T}})$	$\lvert m^{\mathrm{T}}\rvert$	Obj.	CPU
1	2	19	188.37	0	4	34	310.36	0	6	52	401.07	0
2	2	19	188.37	0	4	34	315.20	0	6	52	399.50	0
3	2	20	183.43	0	4	38	314.50	0	6	52	379.79	0
4	3	22	191.72	0	4	39	355.16	0	6	52	392.30	0
5	3	23	218.62	0	5	41	356.09	0	6	52	390.02	0
6	3	24	225.77	0	5	49	405.60	0	6	52	379.59	0
7	2	15	154.13	0	3	22	162.41	0	6	56	391.77	0
8	2	16	166.62	0	3	27	221.69	0	6	56	367.40	0
9	2	17	182.46	0	3	29	232.97	0	6	56	367.18	0
10	2	19	169.69	0	4	33	265.81	0	6	56	381.12	0
11	2	20	179.39	0	4	36	302.32	0	6	56	396.55	0
12	3	21	187.16	0	4	40	326.12	0	6	56	392.43	0
13	1	9	73.00	0	2	18	126.94	0	6	60	318.70	0
14	1	9	60.56	0	3	22	146.06	0	7	61	431.74	0
15	2	12	97.76	0	3	26	181.28	0	7	61	346.78	0
16	2	12	90.08	0	3	30	242.68	0	7	61	333.65	0
17	2	13	98.30	0	4	35	232.81	0	7	61	377.57	0
18	2	14	102.86	0	5	42	265.43	0	7	61	358.55	0
19	1	5	30.66	0	1	9	42.05	0	7	67	343.60	0
20	1	6	42.05	0	2	11	66.66	0	7	67	363.79	0
21	1	7	42.05	0	2	15	95.02	0	7	67	337.02	0
22	1	7	42.05	0	2	19	153.32	0	7	67	404.03	0
23	1	8	42.05	0	3	24	166.04	0	7	67	380.01	0
24	1	9	42.05	0	4	37	244.71	0	7	67	326.20	0
25	1	1	12.51	0	1	2	25.01	0	8	71	423.03	0
26	1	1	12.51	0	1	7	47.81	0	8	71	405.21	0
27	1	1	12.51	0	1	8	47.81	0	8	71	393.37	0
28	1	2	25.01	0	2	11	66.66	0	8	71	479.16	0
29	1	2	25.01	0	3	24	160.49	0	8	71	415.65	0
30	1	7	47.81	0	4	31	204.47	0	8	71	462.46	0
31	0	0	–	–	0	0	–	–	8	72	384.77	0
32	0	0	–	–	1	5	22.80	0	8	72	404.79	0
33	0	0	–	–	1	6	30.40	0	8	72	379.49	0
34	0	0	–	–	2	11	66.66	0	8	72	459.94	0
35	0	0	–	–	3	24	201.80	0	8	72	534.91	0
36	1	5	22.80	0	4	31	218.44	0	8	72	456.97	0

11.4.2 Computational Results

Table 11.4 shows the total cost of the route partitions for the entire week and total CPU times of Algorithm 4 for each of the 108 tactical MMU operation plans based on the 36 test instances. Each row corresponds to the three tactical MMU operation plans that are again based on the three strategic MMU operation plans for the same test instance using formulations (Det-B), (RobΓ-B), and (RobI-B). For each tactical MMU operation plan m^{T}, we specify the number of required MMUs $\nu(m^{\mathrm{T}}) \in \mathbb{N}$ as well as the total number of scheduled sessions $|m^{\mathrm{T}}| \coloneqq \sum_{\boldsymbol{\ell} \in \boldsymbol{L}} m^{\mathrm{T}}_{\boldsymbol{\ell}}$ per week. Note, that some tactical MMU operation plans from

Phase 2 do not schedule any MMU sessions which is indicated by not providing an objective function value and CPU time in Table 11.4.

Analyzing the objective values for the 108 tactical MMU operation plans, we can confirm that a higher number of scheduled sessions tends to lead to longer vehicle routes which is to be expected. However, just as in the previous section, there are some exceptions from this relationship which is not surprising as the vehicle routing is not considered in Phase 2. To ease the understanding of a solution to Phase 3, we illustrate the resulting vehicle routes for the tactical MMU operation plan for Instance 14 with budgeted uncertainty sets in Figure 11.6.

With regard to running times, we observe that we were able to solve all instances of the VRMMU in less than one CPU second which is consistent with fact that the VRMMU is polynomial-time solvable. Recall, that we were not able to deduce a polynomial-time algorithm for the vehicle routing problem with multiple depots. Instead, we presented the integer programming formulation (mVR) which is potentially more computationally challenging to solve in practice. This is however not investigated in the scope of this study.

11.5 Evaluation using SiM-Care

As a final mean of evaluation, we illustrate how MMU operations obtained through P3MMU can be fed into the hybrid agent-based simulation tool SiM-Care that we introduced in Part I. For this purpose, we consider the robust MMU operations for Instance 14 with budgeted uncertainty sets depicted in Figure 11.6. To integrate MMUs into SiM-Care, our implementation of Algorithm 4 generates a set of MMU agents that represents the serviced MMU operation sites. For each operated morning and afternoon session, we define the opening hours of a site to be from 8:30 a.m. to 11:00 a.m. and 1:30 p.m. to 4:00 p.m., respectively. The resulting 4 MMU agents for Instance 14 are then added to the baseline scenario of the considered primary care system that provided the patient demands for this case study; compare Section 11.1. We refer to the resulting scenario as the baseline scenario with MMUs and note that this scenario has 24 primary care physicians – 20 regular physicians (working in the 16 aggregated practices) and 4 MMU sites.

For the evaluation of the baseline scenario with MMUs, we choose the default setup as described in Chapter 5, i.e., 20 independent runs modeling one year preceded by a warm-up of 60 years. Table 11.5 reports the resulting mean performance indicators as well as the associated exact $95\,\%$-confidence intervals for each performance indicator tracked by SiM-Care; compare Section 4.5. Comparing these values to the ones from the baseline scenario without MMUs shown in Table 5.6, we can quantify the impact of the MMU operations in terms of the performance indicators. In particular, we can observe that the additional treatment capacities introduced by the operation of MMUs improve almost all patient and physician indicators. The physicians' expected workload measured through the average number of treatments decreases by $16\,\%$. Due to the increased availability of appointments,

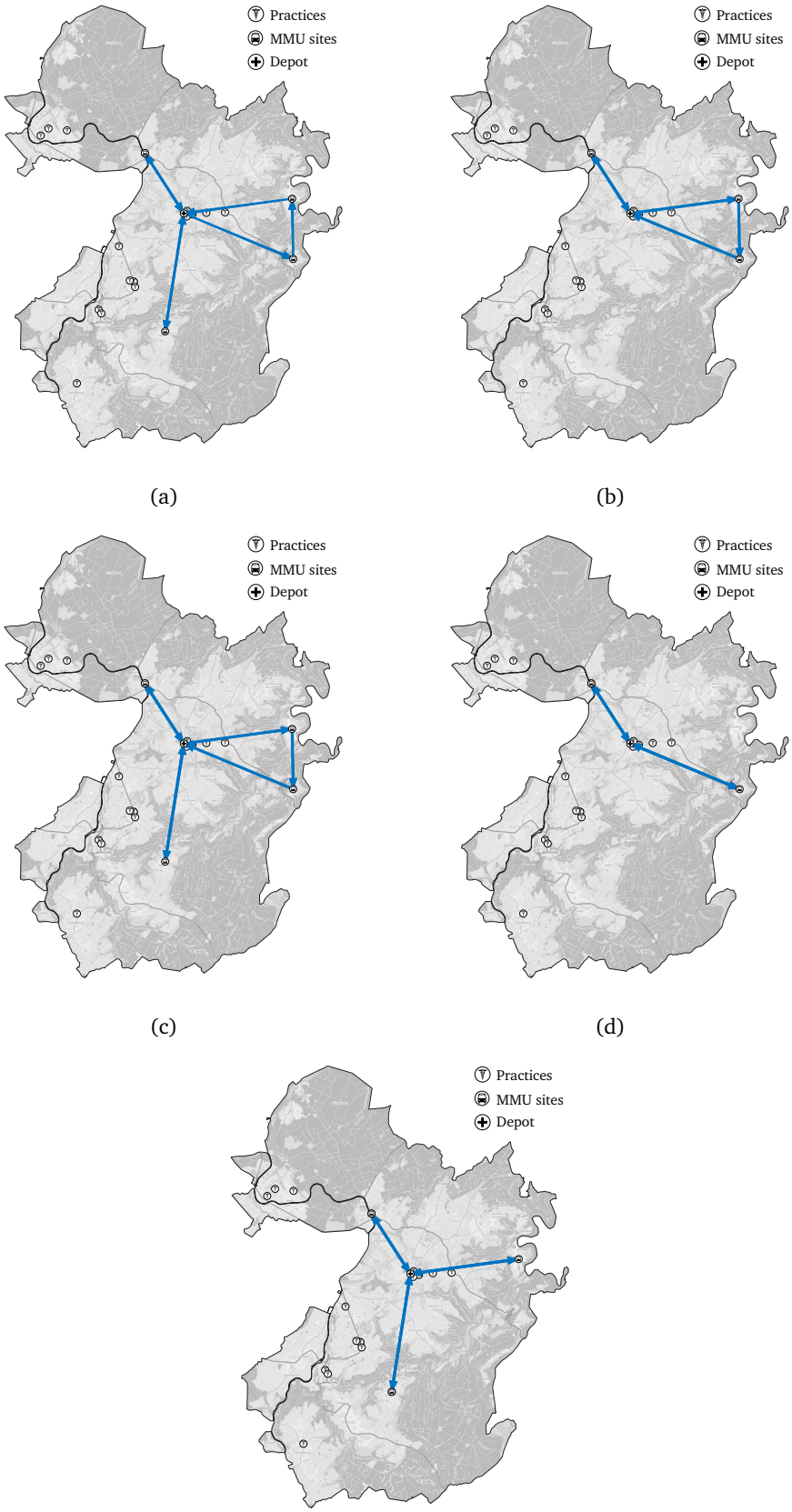


Fig. 11.6.: Route partitions for tactical MMU operation plan based on Instance 14 with budgeted uncertainty sets: (a) Monday, (b) Tuesday, (c) Wednesday, (d) Thursday, and (e) Friday.[7]

Tab. 11.5.: Mean performance indicators and 95 %-confidence intervals obtained by repeating each simulation experiment 20 times for the baseline scenario with MMUs.

	Baseline Scenario with MMUs	
	Mean	95 %-CI
average number of treatments	8,551.38	[8,544.06, 8,558.7]
average number of walk-ins	3,538.53	[3,531.26, 3,545.8]
average number of acute appointments	3,204.03	[3,203.36, 3,204.71]
average number of regular appointments	1,808.81	[1,807.77, 1,809.85]
average utilization [%]	67.46	[67.38, 67.53]
average daily overtime [min]	0.3	[0.27, 0.33]
average number of rejected walk-ins	3.42	[3.12, 3.72]
average access time [d]	2.2	[2.19, 2.2]
average access time regular [d]	1.59	[1.57, 1.61]
average access distance [km]	4.51	[4.5, 4.51]
average waiting time appt. [min]	1.97	[1.95, 1.98]
average waiting time walk-in [min]	34.49	[34.38, 34.6]
on-time appointments [%]	62.79	[62.7, 62.88]
number of acute illnesses	136,363.2	[136,205.55, 136,520.86]
number of chronic patients	10,662	–
total PCP capacity [h]	36,621	–

the average number of patients forced to visit physicians as walk-ins decreases by 25 %. The average daily overtime for physicians (that neglects all the physicians' administrative and organizational tasks) decreases by half a minute. On average, the patients' waiting time for an appointment decreases by 11 %, however, the access time for regular appointments actually increases by almost 7 %. This increase can be explained by the fact that chronic patients who choose an MMU site as their family physician (compare Section 4.2.6) necessarily have to wait longer for their regular appointments as three out of the four MMU sites are only serviced in four sessions of the week. To prevent this behavior, one could adapt the integration of MMUs into SiM-Care such that MMU sites cannot become family physicians. The patients' average access distance decreases by 9 % to 4.5 km which confirms that MMUs can help to overcome access barriers. Finally, the average waiting time for patients with appointment decreases by 6 % while the average waiting time for walk-ins decreases by 13 %.

[7] Map tiles by Humanitarian OSM Team under CC0. Data by OpenStreetMap, under ODbL.

12 Discussion and Conclusion

The aim of P3MMU is to provide an optimization framework for the operational planning of MMUs. We thereby consider a weekly recurring mode of operation with two clinical sessions per day in which MMUs return to their home depot each night. This setting is very rarely considered in the literature and most existing contributions focus on extensive regions in which vehicles return to their depot only once or twice a month.

In Phase 1, we studied the strategic planning problem for MMUs as a capacitated set covering problem. As a new modeling concept, we considered existing infrastructure in the form of practices and both steerable and unsteerable patient demands. While steerable patient demands can be assigned to any acceptable treatment facility, unsteerable demands will always visit the closest available treatment facility. The interplay of the two types of demands brings a new aspect to location planning that has, to the best of our knowledge, not yet been considered in the literature. We formulated the problem as a compact integer linear program and showed how this formulation can be solved by Benders decomposition and constraint generation. Recognizing the importance of uncertainties in health care planning, we extended the problem to uncertain patient demands. Using methods from robust optimization, we devised exact solution methods based on constraint generation for interval and budgeted uncertainty sets.

In Phase 2, we focused on the tactical partitioning problem for MMUs as a partitioning variant of the k-center problem. We presented strong inapproximability results for the problem with both metric and general distance functions before a mathematical programming formulation was introduced. As Phase 1 considers patient demands and treatment capacities in a session-aggregated form that artificially smooths both out over the week, we subsequently introduced a session-specific combined strategic tactical planning problem for MMUs and showed that all our previous results from Phase 1 transfer to it.

In Phase 3, we investigated the vehicle routing problem for MMUs as a special case of the weighted matching problem. In particular, we proved that the problem for a single depot can be reduced to a minimum weight perfect matching problem in a bipartite graph. In the multi-depot setting, we proved that the problem is a special case of the BCBPM which we subsequently showed to be $\mathcal{NP}$-hard.

Our computational study showed that P3MMU enables the computation of optimized MMU operation plans for real-world sized instances in an acceptable time frame. The cost of the MMU operation and thus the cost of the provision of primary care significantly depends on three factors: the percentage of unsteerable patient demands, the modeled consideration sets, and the handling of data uncertainties. Thus, a key finding of this thesis is the insight

that these three factors can have a major impact on the resulting MMU operation plans and should thus be taken into account for strategic MMU operation planning. From our computational experiments, we infer that $(\mathrm{Rob}\Gamma\text{-B})$ may be the most suitable formulation for this purpose. The formulation $(\mathrm{Rob}\Gamma\text{-B})$ allows for the consideration of demand uncertainties while limiting the conservatism of solutions through the use of budgeted uncertainty sets. In addition, it is possible to trade off operational cost against robustness towards demand uncertainties by adjusting the budget parameters Γ_1 and Γ_2.

While this represents a major step forward for MMU operation planning, we must not overlook the limitations of our models and computational study that stem from our assumptions and open up directions for further research. Concerning the limitations of our computational study, let us first note that the assessment of a physician's treatment capacity is a very delicate and personal matter that is beyond our field of expertise. So while our estimates may not be completely unrealistic, we do not make any claims of correctness and would definitely recommend surveying each physician individually. Similar limitations apply to the patient demand origins and corresponding patient demands, where the lack of empirical data forced us to use the simulation model SiM-Care from Part I which can provide rough estimates at best. Moreover, we have seen that the aggregation of patients to demand origins is highly non-trivial and may result in undesired behaviors such as an increase in the total worst-case patient demand. Also note, that our computational experiments neither included the combined strategic tactical planning problem for MMUs nor the vehicle routing with multiple depots. A further more extensive study should therefore investigate whether the respective solution approaches presented in this thesis are computationally tractable for real-world sized instances and determine whether the combined consideration of Phases 1 and 2 improves the solution quality.

With regard to model limitations, we want to note that our assumption that each demand origin's unsteerable demands target the same treatment facility is quite strict and definitely not true in reality. To overcome this limitation, future work should investigate whether this assumption can be weakened, e.g., by assuming that the unsteerable patient demands target the three closest facilities in some fixed ratio. Furthermore, we assume that the patient demands at the demand origins are independent which is questionable in practice. While it could be difficult to remove this assumption entirely from our models, one step into this direction could start by considering steerable and unsteerable patient demands as being dependent. In line with this goal, we could model one joint uncertainty set for the steerable and unsteerable patient demands instead of two separate ones to incorporate their dependencies into our models.

A conceptional weakness of our approach is clearly the loss of optimization potential that results from the separate consideration of the three planning phases. While we presented methods that allow for a combined consideration of the strategic and tactical planning phases, it would be interesting to investigate if one could also include the actual routing of the vehicles into these solution approaches. To that end, one could start by investigating the joint consideration of the tactical planning phase and the vehicle routing as the former has a

high impact on the latter. One way to achieve such an integration could start from a Benders decomposition approach that determines the actual vehicle routes within the subproblem.

Future work on the metric tactical partitioning problem for MMUs (TPMMU) should investigate the existence of a 2-approximation algorithm. Should such a 2-approximation exist, this would imply that our inapproximability result is tight. To come-up with a 2-approximation, one could try to transfer the existing 2-approximation algorithm for the metric k-center problem to the TPMMU. Furthermore, there are various heuristics for the k-center problem, e.g., Mihelič and Robič (2005), which can be potentially adapted to the TPMMU to speed up the solution process of Phase 2 for challenging instances.

Concerning the vehicle routing for MMUs studied in Phase 3, the biggest open challenge is to decide whether the general mVRMMU is $\mathcal{NP}$-hard. The reduction for the BCBPM that we used to prove the strong $\mathcal{NP}$-hardness of several special cases of the mVRMMU could be shown to be unrepresentable for the general mVRMMU. Thus, efforts at proving the problem's $\mathcal{NP}$-hardness have to be based on a different construction. Working towards a polynomial-time algorithm, it might be a good starting point to investigate the polyhedron of the problem formulation (mVR). Alternatively, one could try to adapt algorithms for the minimum cost (multi-commodity) flow problem.

Summing up Part II of this thesis, we are confident that our models produce MMU operation plans that can serve as a sound basis for an actual real-world implementation. That being said, we strongly recommend an expert validation of all plans prior to their implementation due to the discussed limitations. This validation process can be aided by the use of the simulation model SiM-Care as illustrated in our case study.

Part III

Variations of Matching Problems

Multi-Budgeted and Minimum Color-Degree Perfect b-Matchings

13 Introductory Remarks and Contribution

> *Over the past 50 years, many variations on the classic assignment problem have been proposed, a fact that becomes immediately obvious if the key words "assignment problem" are entered into the search engine for the research database ABI/INFORM.*
>
> **— David W. Pentico, 2007**

13.1 Motivation and Research Question

Assignment problems are among the most famous combinatorial optimization problems. In its most basic form, the assignment problem consists of a set of agents A, a set of jobs B, and a set of agent-job pairs $E \subseteq A \times B$ that define which agent can perform which job (Schrijver, 2003); compare Figure 13.1. The objective is to find a one-to-one assignment of jobs to agents. Graph-theoretically the assignment problem corresponds to the maximum (weighted) matching problem in a bipartite graph which is known to be polynomial-time solvable by the Hungarian method (Kuhn, 1955). However, for many applications this original version of the assignment problem fails to capture all relevant requirements. Therefore, various more complex variations of the assignment problem are studied, e.g., the (capacitated) b-matching problem (Schrijver, 2003), the restricted matching problem (Tanimoto et al., 1978), or the stable matching problem (Gale and Shapley, 1962).

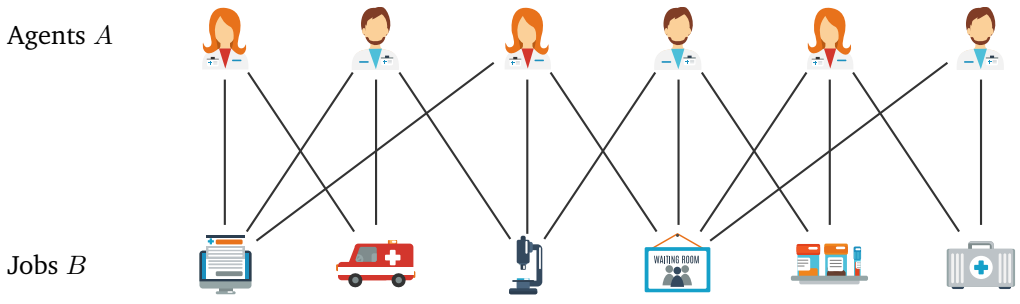


Fig. 13.1.: Example of the assignment problem for the matching of physicians to tasks.[8]

[8] Icon designs by macrovector / Freepik under Freepik terms of use.

In this thesis, we investigate two such variations which arose within our research scope. The first is the *multi-budgeted matching problem* (mBM) which is motivated by the vehicle routing of mobile medical units (MMUs) with multiple depots that we studied in the previous part of this thesis. The second variation is called the *minimum color-degree perfect b-matching problem* (Col-BM) which addresses the assignment of staff to MMU sessions.

While the motivation to consider the mBM and the Col-BM stems from the problems' respective applications in the provision of primary care, we will in the following study both of them from their theoretical side and pose the following question:

> What is each matching problem's complexity and how can we solve it (efficiently) on restricted graph classes?

Before we start answering this question, we introduce both matching problems in more detail and stress their application in the context of this thesis. Then, we summarize our contributions and provide an overview of the related work.

Multi-Budgeted Matchings

The mBM is a budgeted version of the weighted matching problem with k independent edge cost functions. For every cost function, a respective budget limits the total cost of edges we may choose. The objective is to determine a matching of maximum weight such that the cost with respect to each cost function does not exceed the corresponding budget. The mBM generalizes the budgeted colored perfect matching problem (BCBPM) that we encountered in Part II of this thesis to general graphs and multiple cost functions; compare Definition 10.9. Structural results for the mBM can thus be useful to solve the vehicle routing problem for MMUs with multiple depots; see Theorem 10.8. As part of this thesis, we strengthen the strong $\mathcal{NP}$-hardness result for the BCBPM from Theorem 10.10 for the mBM. Moreover, we study the mBM on restricted graph classes and apply dynamic programming techniques to derive new pseudo-polynomial algorithms.

Minimum Color-Degree Perfect b-Matchings

The Col-BM is an extension of the perfect b-matching problem to edge-colored graphs. The objective of the Col-BM is to minimize the maximum number of differently colored edges in a perfect b-matching that are incident to the same node. Given that there are multiple types of MMUs which differ in their internal setup and equipment, the Col-BM can be used to model the staff assignment for MMUs. By representing the vehicle type of each session by the color of edges and matching physicians to MMU sessions, solving the Col-BM corresponds to minimizing the number of different environments physicians have to work in which ultimately improves their efficiency. In this thesis, we show that the Col-BM is strongly $\mathcal{NP}$-hard and

provide an inapproximability result. Then, we derive polynomial-time algorithms for the Col-BM with a fixed number of colors on various graph classes.

13.2 Contribution

The main contributions of Part III of this thesis can be grouped into those concerning the mBM and those addressing the Col-BM. We summarize them in the following.

We show the strong $\mathcal{NP}$-hardness of the mBM on paths with uniform edge weights and budgets by a reduction from 3-SAT. On the algorithmic side, we propose a pseudo-polynomial dynamic program for the mBM with a fixed number of budget constraints on series-parallel graphs which can be extended to trees using a simple graph transformation. Realizing that both these graph classes have a bounded treewidth in common, we come up with a dynamic program based on tree decompositions that solves the mBM with a fixed number of budget constraints on graphs with bounded treewidth in pseudo-polynomial time.

For the Col-BM, we show that the problem is strongly $\mathcal{NP}$-hard on two-colored bipartite graphs by a reduction from (3,B2)-SAT and conclude that there exists no α-approximation algorithm for $1 < \alpha < 2$ unless $\mathcal{P} = \mathcal{NP}$. Algorithmically, we identify a class of two-colored complete bipartite graphs on which we can solve the Col-BM in polynomial time. Furthermore, we use dynamic programming to devise polynomial-time algorithms solving the Col-BM with a fixed number of colors on series-parallel graphs and simple graphs with bounded treewidth.

With these contributions, we extend existing studies of budgeted matching problems and matching problems on edge-colored graphs that we discuss in the following.

13.3 Related Work

Budgeted versions of the weighted matching problem have been previously studied. Berger et al. (2011) consider the budgeted matching problem with a single budget constraint (BM) which is the specialization of the mBM with $k = 1$. A straightforward reduction from the knapsack problem shows the $\mathcal{NP}$-hardness of the BM. The authors introduce a polynomial-time approximation scheme (PTAS) for the BM and show that for polynomial weights and costs, the BM is reducible to the exact perfect matching problem. As a direct consequence of this relationship, the results of Camerini et al. (1992) and Barahona and Pulleyblank (1987) imply a Monte-Carlo pseudo-polynomial algorithm for the BM on general graphs as well as a deterministic pseudo-polynomial algorithm for the BM on planar graphs, respectively.

Previous work on the multi-budgeted matching problem focused primarily on approximation schemes. The first pure approximation scheme is due to Grandoni and Zenklusen (2010).

The authors present a PTAS for 2-budgeted matching, which Chekuri et al. (2011) later generalized to obtain a PTAS for the mBM with a fixed number of budget constraints.

Most edge-colored matching problems impose restrictions that depend on the edge coloring in order to reduce the space of feasible solutions. One of the first problems of this kind is the rainbow or multiple-choice matching problem (Garey and Johnson, 1979): Given an edge-colored graph, find a maximum matching such that all edges have distinct colors. The rainbow matching problem is known to be $\mathcal{NP}$-complete on bipartite graphs (Rusu, 2008), and Le and Pfender (2014) more recently proved that it is even $\mathcal{APX}$-complete on paths. Another problem of this kind is the blue-red matching problem (BRM): Given a blue-red-colored graph and $w \in \mathbb{N}$, find a maximum matching which consists of at most w blue and at most w red edges. Nomikos et al. (2007) devise an $\mathcal{RNC}^2$ as well as an asymptotic $\frac{3}{4}$-approximation algorithm for the BRM. The exact complexity of the BRM is still open.

One of the earliest weighted matching problems considered on edge-colored graphs is the bounded color matching problem (BCM): Given an edge-colored graph with edge weights, find a maximum weighted matching such that the number of edges in each color satisfies a color-specific upper bound. As a generalization of the rainbow matching problem, all complexity results of the former directly translate to the BCM. A straightforward greedy strategy leads to a $\frac{1}{3}$-approximation algorithm for the BCM (Mastrolilli and Stamoulis, 2014). Moreover, several bi-criteria approximation algorithms for the BCM that are allowed to slightly violate the color constraints are due to Mastrolilli and Stamoulis (2012) and Mastrolilli and Stamoulis (2014). Recently, an extension of the BCM that additionally incorporates edge costs was studied under the name budgeted colored matching problem by Büsing and Comis (2018a). All these variations of the BCM are special cases of the mBM in which cost functions are no longer independent, but edges incur (unit-)costs only towards exactly one cost function.

The concept of incorporating an edge-coloring into the objective function of a matching problem is, to our knowledge, relatively new and only few problems of this type have been studied so far. One such problem that is closely related to the Col-BM is the labeled maximum matching problem (LMM): Given an edge-colored graph, the LMM asks for a maximum matching that uses the minimum number of different colors. Monnot (2005) showed that the LMM is $\mathcal{APX}$-complete on bipartite complete graphs and 2-approximable on 2-regular bipartite graphs. Subsequently, Carrabs et al. (2009) presented alternative mathematical formulations and an exact branch-and-bound scheme for the LMM. Another family of related problems are so-called reload cost problems. In reload cost problems, edge colors symbolize different types of transport and costs arise for every change of color at a node. The task is to find a specific subgraph for which the weighted sum of all color changes is minimal. Examples of considered sought-after subgraphs are, e.g., spanning trees (Wirth and Steffan, 2001), paths between two vertices (Gourvès et al., 2009), and tours (Amaldi et al., 2011). For a detailed review of these kinds of problems we refer to Baste et al. (2019).

A weighted b-matching problem with an objective function incorporating the edge coloring is the diverse weighted b-matching problem (D-WBM). The D-WBM can be considered as the counterpart of the Col-BM: Given a weighted edge-colored bipartite graph, the D-WBM asks for a b-matching satisfying upper and lower vertex degree bounds such that the weights of edges incident to the same node are evenly distributed among all colors. In Ahmed et al. (2017), this diversification is ensured by minimizing a quadratic function that penalizes unbalanced weight-color distributions rather than adopting a max-min approach analogous to our min-max approach. The authors also claim D-WBM to be $\mathcal{NP}$-hard, however the key result in their argumentation could not be located in the provided reference.

For a more extensive review on general matching theory we refer to Mastrolilli and Stamoulis (2014) and Pentico (2007).

13.4 Outline and Use of Published Materials

Part III of this thesis is structured as follows. Chapter 14 studies the complexity of the mBM as well as the problem's pseudo-polynomial solvability on various graph classes. Chapter 15 considers the Col-BM and shows its strong $\mathcal{NP}$-hardness before polynomial algorithms on restricted graph classes are derived. Finally, we conclude in Chapter 16 by summarizing our findings and providing directions for future work.

Chapter 14 and parts of Chapters 13 and 16 are based on the publication Büsing and Comis (2018b) and are therefore joint work with my supervisor Christina Büsing. Chapter 15 and parts of Chapters 13 and 16 are based on the publication Anapolska et al. (2021) as well as the preceding conference paper Anapolska et al. (2018) and are therefore joint work with my supervisor Christina Büsing and colleagues Mariia Anapolska and Tabea Krabs.

14 Multi-Budgeted Matching Problems

In this chapter, we study the multi-budgeted matching problem – a weighted matching problem with $k \in \mathbb{N}$ independent edge cost functions. For every cost function, a respective budget limits the total cost of edges we may choose. The objective is to determine a matching of maximum weight such that the cost with respect to each cost function does not exceed the corresponding budget. More formally, this can be restated as follows.

Definition 14.1 (mBM)**.** Let $G = (V, E)$ be a graph and $w : E \to \mathbb{N}$ an edge weight function. Moreover, let $c_i \colon E \to \mathbb{N}$ for $i \in \{1, \dots, k\}$ be a set of cost functions and $B_1, \dots, B_k \in \mathbb{N}$ the corresponding budgets. The *multi-budgeted matching problem* (mBM) asks for a matching $M \subseteq E$ of maximum weight $w(M) = \sum_{e \in M} w(e)$ such that none of the cost functions exceeds its budget, i.e., $\sum_{e \in M} c_i(e) \leq B_i$ for all $i \in \{1, \dots, k\}$.

The mBM is a generalization of the budgeted colored bipartite perfect matching problem (BCBPM) that we introduced in Part II of this thesis to model the vehicle routing problem for MMUs with multiple depots; compare Definition 10.9.

Lemma 14.2. *The mBM generalizes the BCBPM.*

Proof. Let $\mathcal{I}$ be an instance of the BCBPM defined by a bipartite graph $G = (V_A \cup V_B, E)$ with $|V_A| = |V_B|$, edge coloring $E = E_1 \cup \dots \cup E_k$, edge weight function $w : E \to \mathbb{N}$, and budgets $B_i \in \mathbb{N}$ per color class E_i for $i \in \{1, \dots, k\}$. We construct an instance $\mathcal{I}'$ of the mBM on the given graph $G = (V_A \cup V_B, E)$. We set the edge costs to $c'_i(e) = |\{e\} \cap E_i|$ for all $e \in E$ and all $i \in \{1, \dots, k\}$. We choose the budgets $B'_i = B_i$ for $i \in \{1, \dots, k\}$ and set $w'(e) = 2\,|V_A|\,\bar{w} - w(e)$ for all $e \in E$ with $\bar{w} \coloneqq \max\{1, \max_{e \in E} w(e)\}$. We show that $M \subseteq E$ is a budgeted colored perfect matching in $\mathcal{I}$ with weight $w(M)$ if and only if M is a multi-budgeted matching in $\mathcal{I}'$ with weight $w'(M) = 2|V_A|^2\,\bar{w} - w(M)$.

Let $M \subseteq E$ be a budgeted colored perfect matching in $\mathcal{I}$ of weight $w(M)$. By construction, M is feasible for $\mathcal{I}'$ as $\sum_{e \in M} c'_i(e) = |M \cap E_i| \leq B_i = B'_i$ for all $i \in \{1, \dots, k\}$ and has weight

$$w'(M) = \sum_{e \in M} w'(e) = \sum_{e \in M} 2\,|V_A|\,\bar{w} - w(e) = 2\,|V_A|^2\,\bar{w} - w(M).$$

Conversely, let $M \subseteq E$ be a multi-budgeted matching in $\mathcal{I}'$ of weight $w'(M) = 2\,|V_A|^2\,\bar{w} - w(M)$. By construction, M satisfies the budget constraints in $\mathcal{I}$ and it thus suffices to show that M is perfect. Assume the contrary, i.e., $|M| < |V_A|$. Then we have that

$$\begin{aligned} w'(M) &\leq (|V_A| - 1)\,(2\,|V_A|\,\bar{w}) = (2\,|V_A|^2 - 2\,|V_A|)\,\bar{w} < (2\,|V_A|^2 - |V_A|)\,\bar{w} \\ &= 2\,|V_A|^2\,\bar{w} - |V_A|\,\bar{w} \leq 2\,|V_A|^2\bar{w} - w(M) = w'(M) \end{aligned}$$

which is a contradiction. Consequently M is a perfect matching in G and feasible for $\mathcal{I}$. □

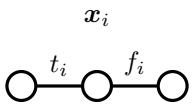
We structure the remainder of this chapter as follows. In Section 14.1, we prove the strong $\mathcal{NP}$-hardness of the mBM on paths. Section 14.2 presents a dynamic program for the mBM with a fixed number of budget constraints on series-parallel graphs and Section 14.3 outlines how this algorithms can be extended to trees via a graph transformation. Finally, Section 14.4 investigates the pseudo-polynomial solvability of the mBM with a fixed number of budget constraints on graphs with bounded treewidth.

14.1 Complexity

Since the mBM generalizes the BCBPM, its strong $\mathcal{NP}$-hardness on bipartite graphs with uniform edge weights and budgets follows immediately from the corresponding strong $\mathcal{NP}$-hardness result of the latter in Theorem 10.10. We strengthen this result by showing that even on paths with uniform edge weights and budgets the mBM is strongly $\mathcal{NP}$-hard. To that end, we reduce 3-SAT to the mBM using a similar construction.

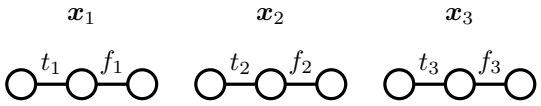
Theorem 14.3. *The decision version of the mBM is strongly $\mathcal{NP}$-complete, even for paths, and uniform edge weights and budgets.*

Proof. Let $\mathcal{I}$ be a 3-SAT instance with n variables $X = \{x_1, \ldots, x_n\}$ and m clauses $C = \{C_1, \ldots, C_m\}$. Every clause $C_j \in C$ has the form $C_j = (y_{j1} \vee y_{j2} \vee y_{j3})$ with $y_{jk} \in L$ for $k \in \{1, 2, 3\}$ and L being the set of literals $L = X \cup \{\bar{x} : x \in X\}$. We construct an instance $\mathcal{I}'$ of the mBM as follows. The graph $G = (V, E)$ is composed of a path of length 2 consisting of the edges $t_i f_i$ for every variable $x_i \in X$. Hence, every maximum matching in G must contain either t_i or f_i for every variable $x_i \in X$ and we associate t_i with the assignment $x_i =$ True and f_i with setting $x_i =$ False; see Figure 14.1(a). We choose unit edge weights and one cost function per clause. For each clause $C_j \in C$ and variable $x_i \in X$, we set $c_j(t_i) = |\{y_{jk} : y_{jk} = \bar{x}_i\}|$ and $c_j(f_i) = |\{y_{jk} : y_{jk} = x_i\}|$. Concerning the available budgets, we set $B_j = 2$ for all $j \in \{1, \ldots, m\}$. By construction, the cost of a matching with respect to cost function c_j indicates the number of unsatisfied literals in clause C_j for all $j \in \{1, \ldots, m\}$. To clarify the previously described construction, it is visualized for an exemplary 3-SAT instance in Figure 14.1(b). We show that $\mathcal{I}$ is a Yes-instance if and only if $\mathcal{I}'$ has a multi-budgeted matching M of weight $w(M) = n$.



$t_i \leftrightarrow x_i = \text{True}$
$f_i \leftrightarrow x_i = \text{False}$

(a)



$C_1 = (x_1 \vee x_1 \vee \bar{x}_3)$ $C_3 = (x_2 \vee x_2 \vee x_3)$
$C_2 = (\bar{x}_1 \vee \bar{x}_1 \vee \bar{x}_3)$ $C_4 = (\bar{x}_2 \vee \bar{x}_2 \vee x_3)$

(b)

	c_1	c_2	c_3	c_4
t_1	0	2	0	0
f_1	2	0	0	0
t_2	0	0	0	2
f_2	0	0	2	0
t_3	1	1	0	0
f_3	0	0	1	1

Fig. 14.1.: (a) Encoding of variables via 2-paths. (b) Visualization of construction via example.

Let $\mathcal{I}$ be a Yes-instance and x^* a satisfying truth assignment. We construct a matching $M^* \subseteq E$ with weight $w(M^*) = n$ as follows: For every $i \in \{1, \dots, n\}$, we pick the edge $t_i \in E$ if $x_i^* = \text{True}$ and $f_i \in E$ otherwise. The resulting matching M^* has obviously weight $w(M^*) = |M^*| = n$ and it hence remains to show that the m budget constraints are satisfied. Assume the contrary, i.e., that there exists a clause $C_j \in C$ such that $\sum_{e \in M^*} c_j(e) = 3$. But this implies that x^* did not satisfy clause C_j which yields a contradiction. The other direction can be shown analogously.

As the decision version of the mBM is obviously in $\mathcal{NP}$ as we can check the weight and feasibility of a given matching in $\mathcal{O}(k \cdot |E|)$ time, the problem's strong $\mathcal{NP}$-completeness follows. We further remark that the collection of 2-paths in our construction can be joined to a single path by edges that induce cost 3 for each budget constraint which ensures their absence in any feasible multi-budgeted matching. □

As soon as we fix the number of budget constraints, such a strong $\mathcal{NP}$-hardness result is not known. Instead, the mBM with a fixed number of budget constraints is only known to be weakly $\mathcal{NP}$-hard; see Grandoni et al. (2014). Hence while there cannot exist a pseudo-polynomial algorithm for the mBM on paths with uniform edge weights and budgets unless $\mathcal{P} = \mathcal{NP}$, there may be pseudo-polynomial algorithms for the mBM with a fixed number of budget constraints. In the following we derive such pseudo-polynomial algorithms for series-parallel graphs, trees, and graphs with bounded treewidth.

14.2 Series-parallel Graphs

We propose a pseudo-polynomial algorithm for the multi-budgeted matching problem with a fixed number of budget constraints on series-parallel graphs and start with a formal definition of series-parallel graphs following the one in Kikuno et al. (1983).

Definition 14.4. A two-terminal graph with distinguished vertices σ and τ called *source* and *sink*, respectively, is called *series-parallel (SP-graph)* if it can be constructed as follows.

1) An edge $\{\sigma, \tau\}$ is series-parallel.

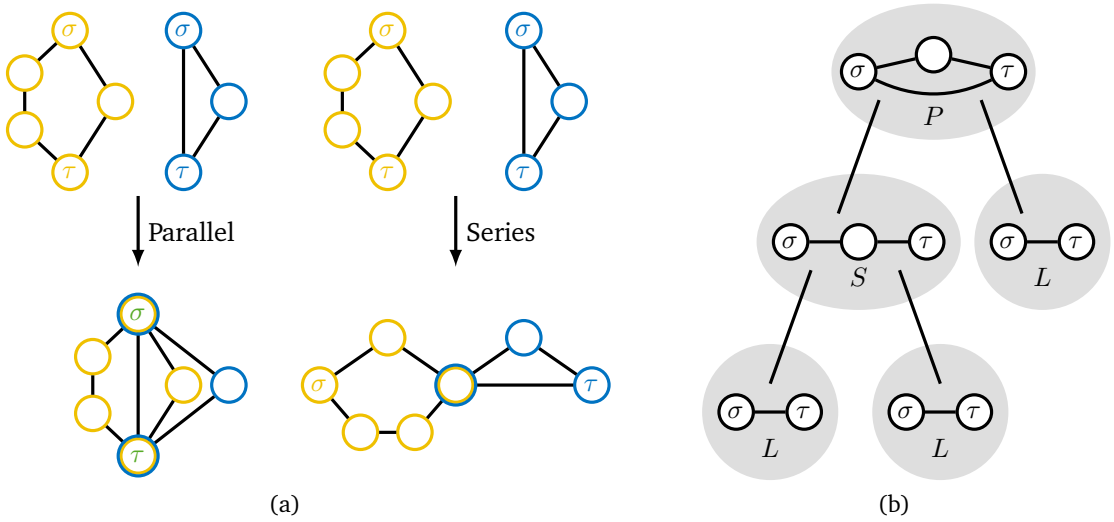


Fig. 14.2.: (a) Parallel and series composition. (b) Example of an SP-tree.

2) A graph constructed by a finite number of the following operations is series-parallel.

   i) Combine two SP-graphs G_1, G_2 with sources σ^1, σ^2 and sinks τ^1, τ^2 by identifying τ^1 with σ^2, called *series composition* of G_1 and G_2.

   ii) Combine two SP-graphs G_1, G_2 with sources σ^1, σ^2 and sinks τ^1, τ^2 by identifying σ^1 with σ^2 and τ^1 with τ^2, called *parallel composition* of G_1 and G_2.

The series and parallel composition are illustrated in Figure 14.2(a). We note that series-parallel graphs may have parallel edges by definition and are generally multi-graphs with designated source and sink vertices. A useful property of SP-graphs is that they can be represented in form of a rooted binary decomposition tree called an *SP-tree*. Given any series-parallel graph G, we can construct an SP-tree $T = T(G)$ that represents the order of series and parallel compositions composing G in polynomial time; see Valdes et al. (1982). The decomposition tree T has three kinds of nodes: *S-nodes*, *P-nodes*, and *L-nodes*. The leaves of T are L-nodes and correspond to the edges of G. The S- and P-nodes are the inner nodes of T and correspond to the subgraph of G obtained by a series- or, respectively, parallel composition of the graphs associated with their two child nodes. We note that the children of S-nodes are ordered because the series composition is not commutative. By construction, the root r of T is associated with the entire graph G. An example of an SP-tree indicating the subgraphs corresponding to the nodes of T is shown in Figure 14.2(b).

Let us now consider the mBM on an SP-graph $G = (V, E)$. We propose a pseudo-polynomial algorithm using dynamic programming on a binary decomposition tree T of G. For a tree node $t \in V(T)$, let G_t denote the subgraph of G with source σ^t and sink τ^t corresponding to t. The dynamic program relies on a set of labels

$$\mathcal{L}^t \coloneqq \{(\alpha, \beta, b) \in \{0,1\} \times \{0,1\} \times \mathbb{N}^k \ : \ b_i \leq B_i \ \forall i \in \{1, \dots, k\}\}$$

defined for every tree node $t \in V(T)$. The parameters $\alpha, \beta \in \{0,1\}$ determine whether the source and sink of G_t are matched or unmatched. The vector $b \in \mathbb{N}^k$ with $b_i \in \{0,\dots,B_i\}$ for $i \in \{1,\dots,k\}$ specifies the available budget with respect to cost function c_i.

For a node $t \in V(T)$ and label $x = (\alpha, \beta, b) \in \mathcal{L}^t$, we call a subset of edges $M \subseteq E(G_t)$ a (t,x)-*restricted matching* if $|\delta_M(\sigma^t)| = \alpha$, $|\delta_M(\tau^t)| = \beta$, $|\delta_M(v)| \leq 1$ for all $v \in V$, and $\sum_{e \in M} c_i(e) \leq b_i$ for all $i \in \{1,\dots,k\}$. Naturally, this leads to the (t,x)-*restricted multi-budgeted matching problem* defined as

$$\max_{M \subseteq E(G_t)} \{w(M) \ : \ M \text{ is } (t,x)\text{-restricted matching in } G_t\}.$$

For a node $t \in V(T)$ and a label $x \in \mathcal{L}^t$, we call the optimal solution value to the (t,x)-restricted mBM the weight $w^t(x)$ of x at t.

We compute the weight of labels based on their children's labels in a bottom-up procedure. If $t \in V(T)$ is an L-node, the corresponding graph G_t consists of a single edge $e \in E$. In case the label $x = (\alpha, \beta, b) \in \mathcal{L}^t$ allows matching both endpoints of e, i.e., $\alpha = \beta = 1$ and provides enough budget to pick e, clearly $M = \{e\}$ is an optimal (t,x)-restricted matching in G_t with weight $w(M) = w(e)$. In all other cases we may not choose e and consequently the optimal obtainable weight is either 0 or $-\infty$ for all infeasible labels. In other words, we can determine the weight of the label $x \in \mathcal{L}^t$ at t as

$$w^t(\alpha,\beta,b) = \begin{cases} w(e) & \text{if } \alpha = \beta = 1,\ b_i \geq c_i(e)\ \forall\, i \in \{1,\dots,k\} \\ -\infty & \text{if } \alpha = \beta = 1,\ \exists\, j \in \{1,\dots,k\} : b_j < c_j(e) \text{ or if } \alpha \neq \beta \\ 0 & \text{else} \end{cases} . \tag{14.1}$$

If $t \in V(T)$ corresponds to the series composition of its two child nodes $\ell \in V(T)$ and $u \in V(T)$, every matching $M^t \subseteq E(G_t)$ in G_t can be decomposed into a matching $M^\ell := M^t \cap E(G_\ell)$ in G_ℓ and a matching $M^u := M^t \cap E(G_u)$ in G_u. We can consequently compute the weight of a label $x = (\alpha, \beta, b) \in \mathcal{L}^t$ at t from the weights of labels at ℓ and u. Combining labels, we have to make sure that the source σ^ℓ of G_ℓ and sink τ^u of G_u are matched as required by the parameters α and β. Moreover, since we contract the sink τ^ℓ of G_ℓ with the source σ^u of G_u we must ensure that the contracted vertex is not matched twice. The budgets b_i for $i \in \{1,\dots,k\}$ must be optimally split between G_ℓ and G_u. Formally, this leads to

$$w^t(\alpha,\beta,b) = \max_{\beta^\ell + \alpha^u \leq 1,\ 0 \leq b' \leq b} \{w^\ell(\alpha,\beta^\ell,b') + w^u(\alpha^u,\beta,b-b')\}. \tag{14.2}$$

Let us now consider $t \in V(T)$ being the parallel composition of $\ell \in V(T)$ and $u \in V(T)$. As for the series composition, we can compute the weight of a label $x = (\alpha, \beta, b) \in \mathcal{L}^t$ from the weights of labels belonging to ℓ and u. Since we contract both the sources and the sinks of G_ℓ and G_u, we must ensure that neither the contracted source $\sigma^t = \sigma^\ell = \sigma^u$, nor the contracted sink $\tau^t = \tau^\ell = \tau^u$ is matched twice. Additionally, the contracted source and sink must be

matched as required by the parameters α and β. The budgets b_i for $i \in \{1, \ldots, k\}$ must be optimally split between G_ℓ and G_u. Formally, this implies

$$w^t(\alpha, \beta, b) = \max_{\substack{\alpha^\ell + \alpha^u = \alpha,\ \beta^\ell + \beta^u = \beta,\\ 0 \leq b' \leq b}} \{w^\ell(\alpha^\ell, \beta^\ell, b') + w^u(\alpha^u, \beta^u, b - b')\}. \tag{14.3}$$

The weight of an optimal multi-budgeted matching in G can now be computed by considering the labels belonging to the root r of T and budgets $b = B$.

Lemma 14.5. *Let $M^* \subseteq E$ denote an optimal multi-budgeted matching in G. Then it holds that*

$$w(M^*) = \max\left\{w^r(\alpha, \beta, B) \ :\ \alpha, \beta \in \{0, 1\}\right\}.$$

Proof. We remind ourselves that by definition $G_r = G$. Therefore, the mBM on G is a relaxation of the (r, x)-restricted mBM for each label $x = (\alpha, \beta, B) \in \mathcal{L}^r$ and thus

$$w(M^*) \geq \max\left\{w^r(\alpha, \beta, B) \ :\ \alpha, \beta \in \{0, 1\}\right\}.$$

For the converse direction, let us define $\alpha^* \coloneqq |\delta_{M^*}(\sigma^r)|$ and $\beta^* \coloneqq |\delta_{M^*}(\tau^r)|$. It is obvious that with these choices M^* is $(r, (\alpha^*, \beta^*, B))$-restricted and it follows that

$$\max\left\{w^r(\alpha, \beta, B) \ :\ \alpha, \beta \in \{0, 1\}\right\} \geq w^r(\alpha^*, \beta^*, B) \geq w(M^*).$$

□

We provide a pseudo-code of our dynamic program in Algorithm 5. An optimal multi-budgeted matching $M^* \subseteq E$ in G can be found by backtracking the chosen maxima in the steps of the dynamic program. The theorem below follows.

Theorem 14.6. *The mBM on SP-graphs parameterized by the number of budget constraints k can be solved in $\mathcal{O}(|E| \cdot \prod_{i=1}^{k} B_i^2)$ time.*

Proof. The correctness of Algorithm 5 directly follows from Lemma 14.5 and the correctness of the label initialization in (14.1) and label updates in (14.2) and (14.3). Analyzing the algorithm's running time, we need to compute the weight of $\mathcal{O}(|E| \cdot \prod_{i=1}^{k} B_i)$ labels as every SP-tree T of $G = (V, E)$ has exactly $|V(T)| = 2\,|E| - 1$ nodes and we have to compute the weight of $4 \prod_{i=1}^{k} (B_i + 1)$ labels for each of them.

It remains to bound the complexity of computing the weight of labels. If $t \in V(T)$ is a leaf, computing label weights is clearly in $\mathcal{O}(1)$ for fixed k. If $t \in V(T)$ corresponds to a series composition, we need to compute the maximum of $3 \prod_{i=1}^{k} (b_i + 1)$ sums which can be done in

Algorithm 5: The mBM on SP-graphs

Input: SP-graph $G = (V, E)$, weights $w : E \to \mathbb{N}$, costs $c_i : E \to \mathbb{N}$ for $i \in \{1, \dots k\}$, budgets $B \in \mathbb{N}^k$

Output: Weight of an optimal multi-budgeted matching in G

compute a binary decomposition tree T of G with root r
let $t_1, \dots, t_n$ be an order on $V(T)$ of non-decreasing height $h(t)$ defined as the number of edges on the longest downward path from t to a leaf of T
for $j = 1, \dots, n$ **do**
  for $(\alpha, \beta, b) \in \mathcal{L}^{t_j}$ **do**
    if t_j *is a leaf in* T *corresponding to edge* $e \in E$ **then**

$$w^{t_j}(\alpha, \beta, b) = \begin{cases} w(e) & \text{if } \alpha = \beta = 1,\ b_i \geq c_i(e)\ \forall\, i \in \{1, \dots, k\} \\ -\infty & \text{if } \alpha = \beta = 1,\ \exists\, j \in \{1, \dots, k\} : b_j < c_j(e) \text{ or if } \alpha \neq \beta \\ 0 & \text{else} \end{cases}$$

    else if t_j *corresponds to the series composition of* $\ell, u \in V(T)$ **then**

$$w^{t_j}(\alpha, \beta, b) = \max_{\beta^\ell + \alpha^u \leq 1,\ 0 \leq b' \leq b} \{w^\ell(\alpha, \beta^\ell, b') + w^u(\alpha^u, \beta, b - b')\}$$

    else if v_j *corresponds to the parallel composition of* $\ell, u \in V(T)$ **then**

$$w^{t_j}(\alpha, \beta, b) = \max_{\substack{\alpha^\ell + \alpha^u = \alpha,\ \beta^\ell + \beta^u = \beta, \\ 0 \leq b' \leq b}} \{w^\ell(\alpha^\ell, \beta^\ell, b') + w^u(\alpha^u, \beta^u, b - b')\}$$

return $\max \{w^r(\alpha, \beta, B)\ :\ \alpha, \beta \in \{0, 1\}\}$

$\mathcal{O}(\prod_{i=1}^k B_i)$ time for fixed k. Analogously, the same bound holds for parallel compositions. Overall, the weight of labels can thus be computed in $\mathcal{O}(\prod_{i=1}^k B_i)$ time.

Combining the bounds on the considered labels and cost of computing their weights, we obtain a total running time in $\mathcal{O}(|E| \cdot \prod_{i=1}^k B_i^2)$. To complete our proof, we remark that Valdes et al. (1982) showed that an SP-tree T of G can be computed in $\mathcal{O}(|E|)$ time. □

As a consequence, Algorithm 5 solves the mBM on series-parallel graphs in pseudo-polynomial time for a fixed number of budget constraints. If we additionally parameterize the mBM by the maximum budget $\bar{B} \coloneqq \max_{1 \leq i \leq k} B_i$ we get the following.

Corollary 14.7. *The mBM on SP-graphs parametrized by the number of budget constraints k and maximum budget $\bar{B}$ is fixed-parameter tractable ($\mathcal{FPT}$).*

14.3 Trees

The second class of graphs for which we present a pseudo-polynomial algorithm for the mBM are trees. Trees are in general not series-parallel as the claw graph $K_{1,3}$ illustrates. We

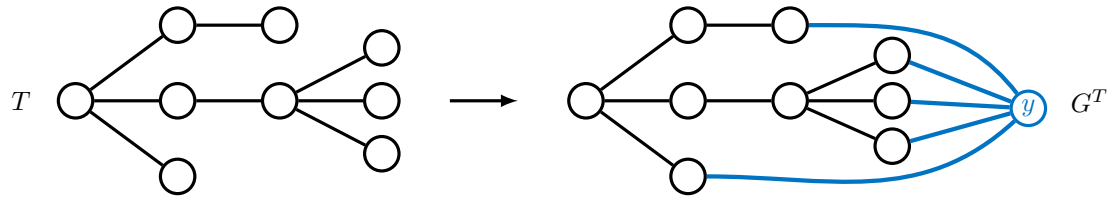


Fig. 14.3.: Construction of series-parallel graph G^T for tree T.

show how trees can be extended to series-parallel graphs by introducing a new node and connecting it to all leaves of the tree.

Definition 14.8. Let $T = (V(T), E(T))$ be a tree. We define G^T as the graph obtained by introducing a new node y to $V(T)$, i.e., $V(G^T) = V(T) \cup \{y\}$ and connecting y to all leaves $L \subseteq V(T)$ of T, i.e., $E(G^T) = E(T) \cup \{\{v, y\} : v \in L\}$.

The construction of G^T as described in Definition 14.8 is visualized in Figure 14.3. We show that for every tree T the graph G^T is series-parallel.

Lemma 14.9. *Let T be a tree. Then G^T is series-parallel.*

Proof (constructive). Let $r \in V(T)$ be an arbitrary node in T. Associate a series-parallel graph $G_v = \{\sigma^v, \tau^v\}$ to every leaf $v \in L$ of T. Let P_v denote the unique path from $v \in V(T)$ to r in T. We define the *depth* of $v \in V(T)$ as $d(v) \coloneqq |P_v|$ and denote the successor of v on P_v by $s(v)$. Consider the nodes $v \in V(T)$ in order of non-increasing depth $d(v)$: Append an edge $\{\sigma, \tau\}$ to G_v through a series composition and associate the resulting SP-graph with the successor $s(v)$ of v. Whenever multiple graphs are associated with the same node, join them via a parallel composition. For the graph G_r associated with r follows $G_r = G^T$. □

Using Lemma 14.9, we can now reduce the mBM on trees to the mBM on SP-graphs.

Theorem 14.10. *The mBM on trees parameterized by the number of budget constraints k can be solved in $\mathcal{O}(|V(T)| \cdot \prod_{i=1}^{k} B_i^2)$ time.*

Proof. Instead of devising new label updating steps exploiting the special structure of trees, we apply the graph transformation from Definition 14.8. Given an mBM instance $\mathcal{I}$ with tree T, budgets $B \in \mathbb{N}^k$, costs $c_i : E(T) \to \mathbb{N}$ for $i \in \{1, \dots, k\}$, and weights $w\colon E(T) \to \mathbb{N}$, we construct an mBM instance $\mathcal{I}'$ on G^T with budgets $B' \in \mathbb{N}^k$, costs $c'_i : E(G^T) \to \mathbb{N}$ for $i \in \{1, \dots, k\}$, and weights $w'\colon E(G^T) \to \mathbb{N}$ as follows. For all edges $e \in E(T)$, we set the weight $w'(e) = w(e)$ and cost $c'_i(e) = c_i(e)$ for $i \in \{1, \dots, k\}$. For all edges $e \in E(G^T) \setminus E(T)$, we set the weight $w'(e) = 0$ and cost $c'_i(e) = \bar{B} + 1$ for $i \in \{1, \dots, k\}$ to ensure their absence in any feasible matching. For $B' = B$, every feasible multi-budgeted matching M in $\mathcal{I}$ is also a feasible multi-budgeted matching in $\mathcal{I}'$ of identical cost and vice versa.

By Lemma 14.9, G^T is series-parallel. The constructed mBM instance $\mathcal{I}'$ can therefore be solved in $\mathcal{O}(|E(G^T)| \cdot \prod_{i=1}^{k} B_i^2)$ time by Theorem 14.6. As $|E(G^T)| \leq |E(T)| + |V(T)| \leq 2\,|V(T)|$ and the construction of $\mathcal{I}'$ can be done in $\mathcal{O}(|V(T)|)$ time the result follows. □

We note that such a transformation is in general not applicable for other budgeted problems, e.g., the budgeted minimum cost flow problem (Büsing et al., 2016).

14.4 Graphs with Bounded Treewidth

This section considers the mBM on graphs of bounded treewidth – a class of graphs that particularly includes SP-graphs and trees. Graphs of bounded treewidth can be decomposed via decomposition trees which enable efficient algorithms for various generally $\mathcal{NP}$-hard problems. To name a few, vertex cover, dominating set, independent set, and Hamiltonian circuit have been shown to be polynomial-time solvable on graphs of bounded treewidth; see Bodlaender (1988). To introduce the concept of treewidth, we start with a formal definition of tree decompositions followed by a definition of the treewidth of a graph according to Robertson and Seymour (1986).

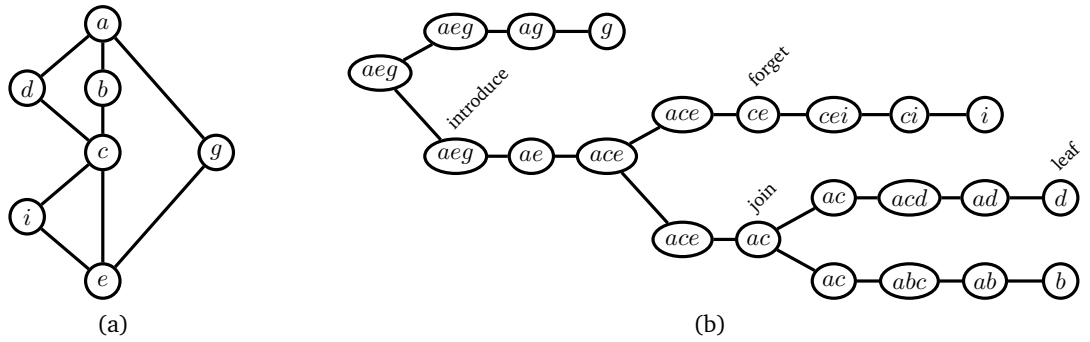
Definition 14.11. For a graph $G = (V, E)$, a pair $(T, \mathcal{X})$ consisting of a tree $T = (V(T), E(T))$ and a collection of vertex subsets (called *bags*) $\mathcal{X} = \{X_t \subseteq V : t \in V(T)\}$ associated to the nodes of T, is called a *tree decomposition* of G if it satisfies the following properties:

1) Every vertex of G is contained in at least one bag, i.e., $\bigcup_{t \in V(T)} X_t = V$.
2) For each edge $\{v, w\} \in E$ there exists a node $t \in V(T)$ such that $v, w \in X_t$.
3) For all nodes $t, \ell, u \in V(T)$ such that ℓ lies on the unique path between t and u in T, it holds that $X_t \cap X_u \subseteq X_\ell$.

The *width* of a tree decomposition of the graph G is defined as the cardinality of its largest bag minus one, i.e., $tw(G, (T, \mathcal{X})) \coloneqq \max_{t \in V(T)} |X_t| - 1$. The *treewidth* of a graph G is then defined as the smallest width among all tree decompositions of G, i.e., $tw(G) \coloneqq \min\{tw(G, (T, \mathcal{X})) : (T, \mathcal{X})$ is a tree decomposition of $G\}$.

Bodlaender (1996) showed that for any given graph G with bounded treewidth, a tree decomposition attaining this width can be constructed in linear time. A special kind of tree decomposition that is regularly used for describing dynamic programs in order to improve their readability are so-called nice tree decompositions (Kloks, 1994; Bodlaender and Koster, 2008).

Definition 14.12. A tree decomposition $(T, \mathcal{X})$ of a graph $G = (V, E)$ is called *nice* if T is a rooted tree and all nodes $t \in V(T)$ can be categorized into four groups:



(a) (b)

Fig. 14.4.: (a) Graph G with $tw(G) = 2$. (b) Nice tree decomposition $(T, \mathcal{X})$ of G.

1) *Leaves* $t \in V(T)$ have no child nodes and their bag contains exactly one vertex $v \in V$, i.e., $X_t = \{v\}$.

2) *Introduce nodes* $t \in V(T)$ have exactly one child node $\ell \in V(T)$ such that $X_\ell \subsetneq X_t$ and $X_t \setminus X_\ell = \{w\}$ for some $w \in V$.

3) *Forget nodes* $t \in V(T)$ have exactly one child node $\ell \in V(T)$ such that $X_t \subsetneq X_\ell$ and $X_\ell \setminus X_t = \{w\}$ for some $w \in V$.

4) *Join nodes* $t \in V(T)$ have exactly two child nodes $\ell, u \in V(T)$ such that $X_t = X_\ell = X_u$.

Any tree decomposition can be transformed into a nice tree decomposition of the same width and $\mathcal{O}(|V|)$ bags in linear time (Kloks, 1994). An example of a graph with bounded treewidth and corresponding nice tree decomposition is depicted in Figure 14.4. We will in the following assume w.l.o.g. that the bag associated with the root r of a nice tree decomposition $(T, \mathcal{X})$ contains only a single node, i.e., $|X_r| = 1$. If this is not the case, we simply add a sequence of $|X_r| - 1$ forget nodes to r and redefine the root of the resulting tree as the last of them.

Let us now consider the mBM on graphs with bounded treewidth. We present a dynamic program for the mBM that recursively updates labels defined on the nodes of a nice tree decomposition. Let $G = (V, E)$ be a graph with bounded treewidth $tw(G) < W$ and $(T, \mathcal{X})$ a corresponding nice tree decomposition such that $|X_t| \leq W$ for all $X_t \in \mathcal{X}$. We denote the set of edges of G induced by X_t by $E[X_t]$ and define G_t to be the subgraph of G induced by the vertices in the bags of the subtree of T rooted in $t \in V(T)$. Thus, from Property 1) of tree decompositions $G_r = G$ directly follows for the root r of T.

We associate a set of labels of the form

$$\mathcal{L}^t = \{(m, o, b) \in \{0,1\}^{E[X_t]} \times \{0,1\}^{X_t} \times \mathbb{N}^k \ : \ b_i \leq B_i \ \forall i \in \{1, \ldots, k\}\}$$

with all nodes $t \in V(T)$. The binary valued mapping $m \colon E[X_t] \to \{0,1\}$ indicates whether the edges in $E[X_t]$ are in the matching or not. The binary valued mapping $o \colon X_t \to \{0,1\}$ indicates whether the vertices in X_t are reserved to be matched to previously forgotten vertices in $V(G_t) \setminus X_t$ or not. Consequently, all vertices $v \in X_t$ with $o_v = 1$ must not be

matched by edges in $E[X_t]$ while vertices with $o_v = 0$ must not be matched by edges in $E(G_t) \setminus E[X_t]$. Finally, the vector $b \in \mathbb{N}^k$ with $b_i \in \{0, \dots, B_i\}$ for $i \in \{1, \dots, k\}$ specifies the available budget per cost function reserved for matching edges in $E(G_t) \setminus E[X_t]$.

We define the weight $w^t(x)$ of label $x = (m, o, b) \in \mathcal{L}^t$ as the maximum weight of a multi-budgeted matching $M \subseteq E(G_t)$ in G_t with the additional constraints implied x. That is, for every $i \in \{1, \dots, k\}$ the available budget for edges in $E(G_t) \setminus E[X_t]$ with respect to cost function c_i is b_i. An edge $e \in E[X_t]$ is part of M if and only if $m_e = 1$ and all vertices $v \in X_t$ with $o_v = 1$ may solely be matched by edges in $E(G_t) \setminus E[X_t]$ while vertices with $o_v = 0$ may only be matched by edges in $E[X_t]$. We call this problem the (t, x)-*restricted multi-budgeted matching problem* and denote it by rmBM(t, x). Hence, the weight $w^t(x)$ of label $x \in \mathcal{L}^t$ at $t \in V(T)$ is the optimal solution value to the rmBM(t, x) which we formally define as

$$\max_{M \subseteq E(G_t)} \quad w(M) \tag{14.4a}$$

$$\text{s.t.} \quad \sum_{e \in M} c_i(e) \leq B_i \qquad \forall i \in \{1, \dots, k\} \tag{14.4b}$$

$$\sum_{e \in M \setminus E[X_t]} c_i(e) \leq b_i \qquad \forall i \in \{1, \dots, k\} \tag{14.4c}$$

$$|e \cap M| = m_e \qquad \forall e \in E[X_t] \tag{14.4d}$$

$$|\delta_M(v) \cap E[X_t]| \leq 1 - o_v \qquad \forall v \in X_t \tag{14.4e}$$

$$|\delta_M(v) \setminus E[X_t]| \leq o_v \qquad \forall v \in X_t \tag{14.4f}$$

$$|\delta_M(v)| \leq 1 \qquad \forall v \in V(G_t). \tag{14.4g}$$

Let $x^t = (m^t, o^t, b^t) \in \mathcal{L}^t$ be a label corresponding to $t \in V(T)$. At the beginning of every label update, we check labels for their validity, i.e., whether rmBM(t, x^t) is feasible. That is, the vertices $v \in X_t$ must not be matched more than once and the cost of the fixed edges within $E[X_t]$ plus the reserved budget b^t must not exceed the total available budget B. Formally this translates as, if there exists $v \in X_t$ such that $o^t_v + \sum_{e \in \delta_{G_t}(v) \cap E[X_t]} m^t_e \geq 2$ or $j \in \{1, \dots, k\}$ such that $b^t_j + \sum_{e \in E[X_t]} c_j(e)\, m^t_e > B_j$, the label is invalid and $w^t(m^t, o^t, b^t) = -\infty$. All remaining labels are called *valid* and we update their weight based on their children's labels in a bottom-up procedure. Before discussing the label updating procedures for the inner nodes of T, we consider the initialization of label weights for the leaves of T.

If $t \in V(T)$ is a leaf, $|X_t| = 1$ and G_t consists of an isolated vertex $v \in V$ such that $E(G_t) = \emptyset$. Consequently, $m^t \colon E[X_t] \to \{0, 1\}$ has an empty domain and we initialize the weight of labels for leaves as follows.

Lemma 14.13. *If $t \in V(T)$ is a leaf of T and $x^t \in \mathcal{L}^t$ we have*

$$w^t(x^t) = 0.$$

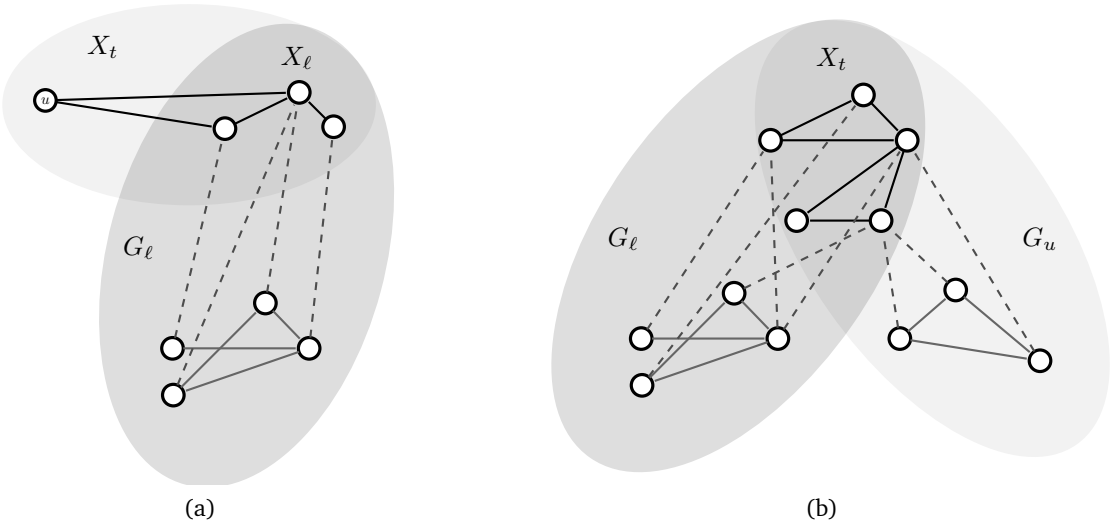


Fig. 14.5.: (a) Visualization of G_t for $t \in V(T)$ being an introduce node with child node $\ell \in V(T)$. (b) Visualization of G_t for $t \in V(T)$ being a join node with child nodes $\ell, u \in V(T)$.

Proof. The graph G_t consists of an isolated vertex and has no edges. Thus, the only feasible and therefore optimal multi-budgeted matching for the rmBM(t, x^t) is the empty matching $M = \emptyset$ of weight $w(M) = 0$. □

If $t \in V(T)$ is an introduce node, we compute the weight of $x^t \in \mathcal{L}^t$ from the previously considered labels of t's child node. Let $\ell \in V(T)$ be the only child of t and $u \in V$ the vertex introduced by t, hence $\{u\} = X_t \setminus X_\ell$. Furthermore, let $U \subseteq E$ be the set of edges introduced by t, i.e., $U \coloneqq E[X_t] \setminus E[X_\ell]$. For a visualization of this setting we refer to Figure 14.5(a).

Lemma 14.14. *Let $t \in V(T)$ be an introduce node with unique child node $\ell \in V(T)$. Given a valid label $x^t = (m^t, o^t, b^t) \in \mathcal{L}^t$, we define the label $x^\ell = (m^\ell, o^\ell, b^\ell) \in \mathcal{L}^\ell$ via $m^\ell_e = m^t_e$ for all $e \in E[X_\ell]$, $o^\ell_v = o^t_v$ for all $v \in X_\ell$, and $b^\ell_i = b^t_i$ for all $i \in \{1, \dots, k\}$. Then we can determine the weight of x^t at t as*

$$w^t(m^t, o^t, b^t) = w^\ell(m^\ell, o^\ell, b^\ell) + \sum_{e \in U} w(e)\, m^t_e.$$

Proof. We show that every feasible matching to the rmBM(t, x^t) can be restricted to a feasible matching to the rmBM(ℓ, x^ℓ). Let $M^t \subseteq E(G_t)$ be an optimal solution to the rmBM(t, x^t), i.e., $w^t(x^t) = w(M^t)$. We define the matching $M^\ell \coloneqq M^t \setminus U$ and show that M^ℓ is feasible for the rmBM(ℓ, x^ℓ), i.e., we show M^ℓ satisfies (14.4b)–(14.4g). For $i \in \{1, \dots, k\}$

$$\sum_{e \in M^\ell} c_i(e) \leq \sum_{e \in M^t} c_i(e) \leq B_i \quad \text{and} \quad \sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) = \sum_{e \in M^t \setminus E[X_t]} c_i(e) \leq b^t_i = b^\ell_i$$

and hence (14.4b) and (14.4c) are satisfied. For $e \in E[X_\ell]$, (14.4d) is satisfied as

$$|e \cap M^\ell| = |e \cap M^t| = m_e^t = m_e^\ell.$$

Concerning (14.4e) and (14.4f), for $v \in X_\ell$ we have that

$$|\delta_{M^\ell}(v) \cap E[X_\ell]| \leq |\delta_{M^t}(v) \cap E[X_t]| \leq 1 - o_v^t = 1 - o_v^\ell$$

and

$$|\delta_{M^\ell}(v) \setminus E[X_\ell]| = |\delta_{M^t}(v) \setminus E[X_t]| \leq o_v^t = o_v^\ell.$$

Finally, as every subset of a matching is a matching, (14.4g) hold which concludes our proof that M^ℓ is feasible for the rmBM(ℓ, x^ℓ) . Hence, we have

$$\begin{aligned} & w^\ell(x^\ell) \geq w(M^\ell) = w(M^t) - \sum_{e \in U} w(e)\, m_e^t = w_t(x^t) - \sum_{e \in U} w(e)\, m_e^t \\ \Leftrightarrow\ & w^t(x^t) \leq w^\ell(x^\ell) + \sum_{e \in U} w(e)\, m_e^t. \end{aligned}$$

Conversely, let M^ℓ be an optimal solution to the rmBM(ℓ, x^ℓ). We define a matching $M^t \coloneqq M^\ell \cup \{e \in U : m_e^t = 1\}$ and show that M^t is feasible for the rmBM(t, x^t). For $i \in \{1, \ldots, k\}$ inequalities (14.4b) and (14.4c) hold as

$$\sum_{e \in M^t \setminus E[X_t]} c_i(e) = \sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) \leq b_i^\ell = b_i^t$$

which, in combination with the validity of x^t, directly implies that

$$\sum_{e \in M^t} c_i(e) = \sum_{e \in M^t \cap E[X_t]} c_i(e) + \sum_{e \in M^t \setminus E[X_t]} c_i(e) \leq \sum_{e \in E[X_t]} c_i(e)\, m_e^t + b_i^t \leq B_i.$$

By construction, $M^t \cap E[X_t] = \{e \in E[X_t] : m_e^t = 1\}$ and consequently equations (14.4d) hold. As x^t is valid, (14.4g) hold and we can show that (14.4e) holds for $v \in X_t$ as

$$|\delta_{M^t}(v) \cap E[X_t]| = \sum_{e \in \delta_{G_t}(v) \cap E[X_t]} m_e^t \leq 1 - o_v^t.$$

Concerning (14.4f) we have to distinguish cases. For the introduced vertex $u \in X_t$ inequality (14.4f) is trivially satisfied, as $|\delta_{G_t}(u) \setminus E[X_t]| = 0$. For the remaining vertices $v \in X_\ell$

$$|\delta_{M^t}(v) \setminus E[X_t]| = |\delta_{M^\ell}(v) \setminus E[X_\ell]| \leq o_v^\ell = o_v^t.$$

As a result, M^t is feasible for the rmBM(t, x^t) and it follows that

$$\begin{aligned} w^t(x^t) \geq w(M^t) &= w(M^\ell) + \sum_{e \in U} w(e)\, m_e^t \\ &= w^\ell(x^\ell) + \sum_{e \in U} w(e)\, m_e^t \end{aligned}$$

which completes our proof. □

Next, let $t \in V(T)$ be a forget node. We consider the only child $\ell \in V(T)$ of t and let $u \in X_\ell$ be the vertex t forgets, i.e., $\{u\} = X_\ell \setminus X_t$. Moreover, we set $U \subseteq E$ to be the edges that were in the bag of ℓ and have been forgotten by t, i.e., $U \coloneqq E[X_\ell] \setminus E[X_t]$. Thus $G_t = G_\ell$ but decisions for edges $e \in U$ are no longer fixed by labels.

Lemma 14.15. *Let $t \in V(T)$ be a forget node with unique child node $\ell \in V(T)$ and forgotten vertex $\{u\} = X_\ell \setminus X_t$. Given a valid label $x^t = (m^t, o^t, b^t) \in \mathcal{L}^t$, we define a set of labels $\mathcal{L}^\ell(x^t) \subseteq \mathcal{L}^\ell$ at ℓ as*

$$\begin{aligned} \mathcal{L}^\ell(x^t) \coloneqq \Big\{ (m^\ell, o^\ell, b^\ell) \in \mathcal{L}^\ell \;:\; & m_e^\ell = m_e^t && \forall e \in E[X_t], \\ & o_v^\ell = o_v^t - m_{\{v,u\}}^\ell && \forall v \in X_t, \\ & b_i^\ell = b_i^t - \sum_{e \in U} c_i(e)\, m_e^\ell && \forall i \in \{1, \dots, k\} \Big\} \end{aligned}$$

where we define $m_{\{v,u\}}^\ell \coloneqq 0$ for all $\{v, u\} \notin E$ to ease notation. Then we can compute the weight of the label x^t at t as

$$w^t(x^t) = \max_{x^\ell \in \mathcal{L}^\ell(x^t)} w^\ell(x^\ell).$$

Proof. We show that every feasible matching for the rmBM(t, x^t) is also feasible for the rmBM(ℓ, x^ℓ) for some label $x^\ell \in \mathcal{L}^\ell(x^t)$. Let $M^t \subseteq E(G_t)$ be an optimal solution to the rmBM(t, x^t). We construct a label $y^\ell = (m^\ell, o^\ell, b^\ell) \in \mathcal{L}^\ell$ by setting

$$m_e^\ell = \begin{cases} m_e^t & \text{for } e \in E[X_t] \\ |e \cap M^t| & \text{for } e \in U \end{cases}$$

$$o_v^\ell = \begin{cases} o_v^t - m_{\{v,u\}}^\ell & \text{for } v \in X_t \\ 1 - |\delta_{M^t}(v) \cap E[X_\ell]| & \text{for } v = u \end{cases}$$

$$b_i^\ell = b_i^t - \sum_{e \in U} c_i(e)\, m_e^\ell.$$

Obviously, $y^\ell \in \mathcal{L}^\ell(x^t)$ and we show that M^t is feasible for the rmBM(ℓ, y^ℓ). Inequalities (14.4b) and (14.4g) are satisfied as M^t is feasible for the rmBM(t, x^t). Concerning constraints (14.4c), for all $i \in \{1, \dots, k\}$ it holds that

$$\begin{aligned}\sum_{e \in M^t \setminus E[X_\ell]} c_i(e) &= \sum_{e \in M^t \setminus E[X_t]} c_i(e) - \sum_{e \in U} c_i(e)\, m_e^\ell \\ &\leq b_i^t - \sum_{e \in U} c_i(e)\, m_e^\ell = b_i^\ell.\end{aligned}$$

By definition, $|e \cap M^t| = m_e^\ell$ for $e \in U$. For $e \in E[X_\ell] \setminus U$

$$|e \cap M^t| = m_e^t = m_e^\ell$$

and (14.4d) is satisfied. Concerning the forgotten vertex $u \in X_\ell$, $|\delta_{M^t}(u) \cap E[X_\ell]| = 1 - o_u^\ell$ holds by definition. For the remaining vertices $v \in X_\ell \setminus \{u\}$ (14.4e) is also satisfied as

$$|\delta_{M^t}(v) \cap E[X_\ell]| = |\delta_{M^t}(v) \cap E[X_t]| + m_{\{v,u\}}^\ell \leq 1 - o_v^t + m_{\{v,u\}}^\ell = 1 - o_v^\ell.$$

To show (14.4f) we distinguish cases. For all vertices $v \in X_\ell \setminus \{u\}$ we have

$$|\delta_{M^t}(v) \setminus E[X_\ell]| = |\delta_{M^t}(v) \setminus E[X_t]| - m_{\{v,u\}}^\ell \leq o_v^t - m_{\{v,u\}}^\ell = o_v^\ell$$

whereas for node $u \in X_\ell$ we can make use of the definition of o_u^ℓ to show

$$\begin{aligned}&1 \geq |\delta_{M^t}(u)| = |\delta_{M^t}(u) \setminus E[X_\ell]| + |\delta_{M^t}(u) \cap E[X_\ell]| \\ \Leftrightarrow\ &1 - |\delta_{M^t}(u) \cap E[X_\ell]| \geq |\delta_{M^t}(u) \setminus E[X_\ell]| \\ \Leftrightarrow\ &o_u^\ell \geq |\delta_{M^t}(u) \setminus E[X_\ell]|.\end{aligned}$$

Consequently, M^t is a feasible matching for the rmBM(ℓ, y^ℓ) which implies

$$w^\ell(y^\ell) \geq w(M^t) = w^t(x^t)$$

and as $y^\ell \in \mathcal{L}^\ell(x^t)$ it directly follows that

$$w^t(x^t) \leq w^\ell(y^\ell) \leq \max_{x^\ell \in \mathcal{L}^\ell(x^t)} w^\ell(x^\ell).$$

For the converse direction, it suffices to realize that for every label $x^\ell \in \mathcal{L}^\ell(x^t)$, the rmBM$(t, x^t)$ is a relaxation of the rmBM(ℓ, x^ℓ) and therefore

$$w^t(x^t) \geq \max_{x^\ell \in \mathcal{L}^\ell(x^t)} w^\ell(x^\ell).$$

□

Finally, let $t \in V(T)$ be a join node with children $\ell, u \in V(T)$ and recall that $X_t = X_\ell = X_u$ but generally $G_t \neq G_\ell \neq G_u$. Instead, the graphs G_ℓ and G_u are subgraphs of G_t with $G_t = G_\ell \cup G_u$ such that $V(G_\ell) \cap V(G_u) = X_t$ and $E(G_\ell) \cap E(G_u) = E[X_t]$; refer to Figure 14.5(b). We determine the weight of labels $x^t \in \mathcal{L}^t$ by combining suitable labels $x^\ell \in \mathcal{L}^\ell$ and $x^u \in \mathcal{L}^u$. To that end, we define the set $\mathcal{L}^{\ell,u}(x^t)$ of label pairs $(x^\ell, x^u) \in \mathcal{L}^\ell \times \mathcal{L}^u$ as

$$\mathcal{L}^{\ell,u}(x^t) \coloneqq \Big\{((m^\ell,o^\ell,b^\ell),(m^u,o^u,b^u)) \in \mathcal{L}^\ell \times \mathcal{L}^u \;:\; m^\ell_e = m^u_e = m^t_e \qquad \forall e \in E[X_t], \tag{14.5}$$

$$o^\ell_v + o^u_v = o^t_v \qquad \forall v \in X_t, \tag{14.6}$$

$$b^\ell_i + b^u_i = b^t_i \qquad \forall i \in \{1,\dots,k\}\Big\}. \tag{14.7}$$

Based on this set of labels, we can compute the weight of labels for join nodes.

Lemma 14.16. *Let $t \in V(T)$ be a join node and $x^t \in \mathcal{L}^t$ a valid label at t. Then for each $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$ holds*

$$w^t(x^t) \geq w^\ell(x^\ell) + w^u(x^u) - \sum_{e \in E[X_t]} w(e)\, m^t_e.$$

Proof. Let $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$, $M^\ell \subseteq E(G_\ell)$ be an optimal solution to the rmBM(ℓ, x^ℓ), and $M^u \subseteq E(G_u)$ be an optimal solution to the rmBM(u, x^u). We set $M^t \coloneqq M^\ell \cup M^u$ and show that M^t is a feasible matching for the rmBM(t, x^t). For $i \in \{1,\dots,k\}$, we use the validity of x^t to show (14.4b), as

$$\begin{aligned}\sum_{e \in M^t} c_i(e) &= \sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) + \sum_{e \in M^u \setminus E[X_u]} c_i(e) + \sum_{e \in M^t \cap E[X_t]} c_i(e) \\ &\leq b^\ell_i + b^u_i + \sum_{e \in E[X_t]} c_i(e)\, m^t_e = b^t_i + \sum_{e \in E[X_t]} c_i(e)\, m^t_e \leq B_i.\end{aligned}$$

Additionally, inequalities (14.4c) are satisfied as we have

$$\sum_{e \in M^t \setminus E[X_t]} c_i(e) = \sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) + \sum_{e \in M^u \setminus E[X_u]} c_i(e) \leq b^\ell_i + b^u_i = b^t_i.$$

For all $e \in E[X_t]$ equation (14.4d) holds, since

$$|e \cap M^t| = |e \cap M^\ell| = m^\ell_e = m^t_e.$$

To show (14.4e) and (14.4f), let $v \in X_t$. As o^ℓ_v, o^u_v, and o^t_v are binary, equation (14.6) implies that either $o^\ell_v = o^t_v$ or $o^u_v = o^t_v$. Let us assume w.l.o.g. that $o^\ell_v = o^t_v$, then we have

$$|\delta_{M^t}(v) \cap E[X_t]| = |\delta_{M^\ell}(v) \cap E[X_\ell]| \leq 1 - o^\ell_v = 1 - o^t_v$$

and

$$|\delta_{M^t}(v)\backslash E[X_t]| = |\delta_{M^\ell}(v)\backslash E[X_\ell]| + |\delta_{M^u}(v)\backslash E[X_u]| \leq o_v^\ell + o_v^u = o_v^t.$$

As a result, M^t is feasible for the rmBM$(t, x^t,)$ and it follows that

$$\begin{aligned} w^t(x^t) \geq w(M^t) &= w(M^\ell) + w(M^u) - \sum_{e \in E[X_t] \cap M^t} w(e) \\ &= w^\ell(x^\ell) + w^u(x^u) - \sum_{e \in E[X_t]} w(e)\, m_e^t. \end{aligned}$$

Note, that we subtract $\sum_{e \in E[X_t]} w(e)\, m_e^t$ in the equation above as we would otherwise count the weight of edges in the intersection $M^\ell \cap M^u = M^t \cap E[X_t]$ twice, once in M^ℓ and once in M^u. □

To complete the computation of label costs for join nodes, it remains to show the following.

Lemma 14.17. *Let $t \in V(T)$ be a join node and $x^t \in \mathcal{L}^t$ a valid label at t. Then there always exists a pair of labels $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$ such that*

$$w^t(x^t) \leq w^\ell(x^\ell) + w^u(x^u) - \sum_{e \in E[X_t]} w(e)\, m_e^t.$$

Proof. We show that every feasible matching to the rmBM(t, x^t) can be restricted to a feasible matching of the rmBM(ℓ, x^ℓ) and the rmBM(u, x^u) for some $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$. Let $M^t \subseteq E(G_t)$ be an optimal solution to the rmBM(t, x^t). We define the label $x^\ell = (m^\ell, o^\ell, b^\ell) \in \mathcal{L}^\ell$ as

$$m_e^\ell \coloneqq m_e^t \qquad \forall e \in E[X_\ell] \tag{14.8}$$

$$o_v^\ell \coloneqq |(\delta_{G_\ell}(v) \setminus E[X_\ell]) \cap M^t| \qquad \forall v \in X_\ell \tag{14.9}$$

$$b_i^\ell \coloneqq \sum_{e \in M^t \cap E(G_\ell) \setminus E[X_\ell]} c_i(e) \qquad \forall i \in \{1, \ldots, k\} \tag{14.10}$$

and the label $x^u = (m^u, o^u, b^u) \in \mathcal{L}^u$ as

$$m_e^u \coloneqq m_e^t \qquad \forall e \in E[X_u] \tag{14.11}$$

$$o_v^u \coloneqq o_v^t - o_v^\ell \qquad \forall v \in X_u \tag{14.12}$$

$$b_i^u \coloneqq b_i^t - b_i^\ell \qquad \forall i \in \{1, \ldots, k\}. \tag{14.13}$$

By definition, it holds that $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$. We define the two matchings $M^\ell \coloneqq M^t \cap E(G_\ell)$ and $M^u \coloneqq M^t \cap E(G_u)$ and show that M^ℓ and M^u are feasible for the rmBM(ℓ, x^ℓ) and the rmBM(u, x^u), respectively. Let us start by considering M^ℓ. Since $M^\ell \subseteq M^t$, inequalities

(14.4b) and (14.4g) are clearly satisfied. Moreover, by (14.10) and the definition of M^ℓ we have that

$$\sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) = b_i^\ell$$

for all $i \in \{1, \dots, k\}$ which proves (14.4c). Equations (14.4d) hold as for all $e \in E[X_\ell]$

$$|e \cap M^\ell| = |e \cap M^t| = m_e^t = m_e^\ell.$$

For all vertices $v \in X_\ell$, we can conclude from (14.6) that $o_v^t \geq o_v^\ell$. Therefore,

$$|\delta_{M^\ell}(v) \cap E[X_\ell]| = |\delta_{M^t}(v) \cap E[X_t]| \leq 1 - o_v^t \leq 1 - o_v^\ell$$

which shows (14.4e). Furthermore, we can use (14.9) to show that (14.4f) hold, since

$$|\delta_{M^\ell}(v) \setminus E[X_\ell]| = |(\delta_{G_\ell}(v) \setminus E[X_\ell]) \cap M^t| = o_v^\ell.$$

Hence, we can conclude that M^ℓ is feasible for the rmBM(ℓ, x^ℓ). It remains to show that M^u is feasible for the rmBM(u, x^u). Again, inequalities (14.4b) and (14.4g) hold as $M^u \subseteq M^t$. We can use (14.10) to show that for $i \in \{1, \dots, k\}$ inequality (14.4c) holds, as

$$\sum_{e \in M^u \setminus E[X_u]} c_i(e) = \sum_{e \in M^t \setminus E[X_t]} c_i(e) - \sum_{e \in M^\ell \setminus E[X_\ell]} c_i(e) \leq b_i^t - b_i^\ell = b_i^u.$$

For edges $e \in E[X_u]$ equation (14.4d) holds, since

$$|e \cap M^u| = |e \cap M^t| = m_e^t = m_e^u.$$

For vertices $v \in X_u$, we note that (14.6) also implies that $o_v^t \geq o_v^u$. Therefore,

$$|\delta_{M^u}(v) \cap E[X_u]| = |\delta_{M^t}(v) \cap E[X_t]| \leq 1 - o_v^t \leq 1 - o_v^u$$

and in combination with (14.9) it follows that

$$|\delta_{M^u}(v) \setminus E[X_u]| = |\delta_{M^t}(v) \setminus E[X_t]| - |\delta_{M^\ell}(v) \setminus E[X_\ell]| \leq o_v^t - o_v^\ell = o_v^u$$

which shows (14.4e) and (14.4f). Thus, M^u is feasible for the rmBM(u, x^u) and the lemma's statement follows from the definitions of M^ℓ and M^u, i.e.,

$$\begin{aligned} w^t(x^t) = w(M^t) &= w(M^\ell) + w(M^u) - \sum_{e \in E[X_t] \cap M^t} w(e) \\ &\leq w^\ell(x^\ell) + w^u(x^u) - \sum_{e \in E[X_t]} w(e)\, m_e^t. \end{aligned}$$

□

We can now combine the statements of Lemmas 14.16 and 14.17 into the following corollary describing the label updating procedure for join nodes.

Corollary 14.18. *Let $t \in V(T)$ be a join node and $x^t \in \mathcal{L}^t$ a valid label at t. The weight of x^t at t can be computes as*

$$w^t(x^t) = \max_{(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)} w^\ell(x^\ell) + w^u(x^u) - \sum_{e \in E[X_t]} w(e)\, m_e^t.$$

This concludes our description of the label updating procedures. It remains to clarify how a solution to the mBM on G can be retrieved from the computed label weights.

Lemma 14.19. *Let r be T's root with $X_r = \{z\}$ and $M^* \subseteq E$ an optimal multi-budgeted matching in G. We define the set $\mathcal{L}^* = \{(m, o, B) \in \mathcal{L}^r \ : \ o_z = 1\}$ of labels at r. Then*

$$w(M^*) = \max_{x \in \mathcal{L}^*} w^r(x).$$

Proof. We recall that $G_r = G$ and $E[X_r] = \emptyset$ which implies that actually $|\mathcal{L}^*| = 1$. For the rmBM(r, x) with $x \in \mathcal{L}^*$, inequalities (14.4c) are redundant to (14.4b), while equations (14.4d) are not present. Moreover, inequalities (14.4e), (14.4f), and (14.4g) are redundant as they are trivially satisfied by every matching. Thus, the rmBM(r, x) is equivalent to the mBM on G and it holds that $w^r(x) = w(M^*)$ for $x \in \mathcal{L}^*$. □

A pseudo-code for our dynamic program is provided in Algorithm 6. An optimal multi-budgeted matching $M^* \subseteq E$ can be found by backtracking the chosen maxima in the steps of the dynamic program. We show that Algorithm 6 solves the mBM with a fixed number of budgets constraints on graphs with bounded treewidth in pseudo-polynomial time.

Theorem 14.20. *Let $G = (V, E)$ be a simple graph with bounded treewidth $tw(G) < W$. Then the mBM with fixed k on G can be solved in $\mathcal{O}(W^2\, 2^{W^2+W}\, |V|\, \prod_{i=1}^k B_i^2)$ time.*

Proof. The correctness of Algorithm 6 directly follows from Lemma 14.19 and the correctness of the label updates in Lemmas 14.13, 14.14, 14.15, and Corollary 14.18. We analyze the algorithm's running time in two steps. First, we bound the number of label weights we need to compute and then we bound the complexity of computing the weight of labels.

Recall that a nice tree decomposition $(T, \mathcal{X})$ of G has $\mathcal{O}(|V|)$ nodes. For each node $t \in V(T)$, we have to consider all labels $(m^t, o^t, b^t) \in \mathcal{L}^t$. As G is simple, $|E[X_t]| \leq |X_t|\,(|X_t| - 1) \leq W^2 - W$ which results in $\mathcal{O}(2^{W^2-W})$ possible combinations for m^t. For the domain of the binary valued function o^t holds that $|X_t| \leq W$ and hence there are at most $\mathcal{O}(2^W)$ combinations. The vector $b^t \in \mathbb{N}^k$ with $b_i^t \in \{0, \ldots, B_i\}$ for all $i \in \{1, \ldots, k\}$ determines the reserved budget per cost function. Consequently, we have to consider $\prod_{i=1}^k (B_i + 1)$

Algorithm 6: The mBM on graphs with bounded treewidth

Input: Graph $G = (V, E)$ with bounded treewidth $tw(G) < W$, weights $w : E \to \mathbb{N}$, cost functions $c_i : E \to \mathbb{N}$ for $i \in \{1, \dots, k\}$, budgets $B_1, \dots, B_k \in \mathbb{N}$

Output: Weight of an optimal multi-budgeted matching in G

compute a nice tree decomposition $(T, \mathcal{X})$ of G

let $t_1, \dots, t_n$ be an order on $V(T)$ of non-decreasing height $h(t)$ defined as the number of edges on the longest downward path from t to a leaf

for $j = 1, \dots, n$ **do**

  for $x^{t_j} = (m^{t_j}, o^{t_j}, b^{t_j}) \in \mathcal{L}^{t_j}$ **do**

    if *t_j is leaf in T* **then**

$$w^{t_j}(m^{t_j}, o^{t_j}, b^{t_j}) = 0$$

    else if *there exists $v \in X_{t_j}$ with $o_v^{t_j} + \sum_{e \in \delta_G(v) \cap E[X_{t_j}]} m_e^{t_j} \geq 2$ or $i \in \{1, \dots, k\}$ such that $b_i^{t_j} + \sum_{e \in E[X_{t_j}]} c_i(e)\, m_e^{t_j} > B_i$* **then**

$$w^{t_j}(m^{t_j}, o^{t_j}, b^{t_j}) = -\infty \qquad \texttt{//label is invalid}$$

    else if *t_j is an introduce node in T with introduced edges $U = E[X_t] \setminus E[\ell]$ and ℓ the unique child of t_j in T* **then**

$$w^{t_j}(m^{t_j}, o^{t_j}, b^{t_j}) = w^{\ell}(m^{\ell}, o^{\ell}, b^{t_j}) + \sum_{e \in U} w(e)\, m_e^{t_j}$$

      where $m_e^{\ell} = m_e^{t_j}$ for all $e \in E[X_\ell]$ and $o_v^{\ell} = o_v^{t_j}$ for all $v \in X_\ell$

    else if *t_j is a forget node in T with forgotten vertex $\{u\} = X_\ell \setminus X_t$, forgotten edges $U = E[X_\ell] \setminus E[X_t]$, and ℓ the only child of t_j in T* **then**

$$w^{t_j}(m^{t_j}, o^{t_j}, b^{t_j}) = \max_{x^{\ell} \in \mathcal{L}^{\ell}(x^{t_j})} w^{\ell}(x^{\ell})$$

$$\text{where } \mathcal{L}^{\ell}(x^{t_j}) = \Big\{ x^{\ell} \in \mathcal{L}^{\ell} : m_e^{\ell} = m_e^{t_j} \quad \forall e \in E[X_{t_j}],\; o_v^{\ell} = o_v^{t_j} - m_{\{v,u\}}^{\ell} \quad \forall v \in X_{t_j},\; b_i^{\ell} = b_i^{t_j} - \sum_{e \in U} c_i(e)\, m_e^{\ell} \quad \forall i \in \{1, \dots, k\} \Big\}$$

    else if *t_j is a join node in T and ℓ, u the children of t_j in T* **then**

$$w^{t_j}(m^{t_j}, o^{t_j}, b^{t_j}) = \max_{\{x^{\ell}, x^{u}\} \in \mathcal{L}^{\ell,u}(x^{t_j})} w^{\ell}(x_\ell) + w^{u}(x_u) - \sum_{e \in E[X_{t_j}]} w(e)\, m_e^{t_j}.$$

$$\text{where } \mathcal{L}^{\ell,u}(x^{t_j}) = \Big\{ (x^{\ell}, x^{u}) \in \mathcal{L}^{\ell} \times \mathcal{L}^{u} : m_e^{\ell} = m_e^{u} = m_e^{t_j} \quad \forall e \in E[X_{t_j}],\; o_v^{\ell} + o_v^{u} = o_v^{t_j} \quad \forall v \in X_{t_j},\; b_i^{\ell} + b_i^{u} = b_i^{t_j} \quad \forall i \in \{1, \dots, k\} \Big\}$$

return $\max_{x \in \mathcal{L}^r} w^r(x)$

different values for b^t. Putting everything together, we have to compute the weight for $\mathcal{O}(|V|\, 2^{W^2} \prod_{i=1}^{k} B_i)$ labels.

The initial validity check of labels can be performed in $\mathcal{O}(kW^2)$ time. Computing the weight of labels for leaves can clearly be done in constant time. For introduce nodes, we can compute the weight of labels in $\mathcal{O}(W)$ time. Regarding the weight of labels corresponding to forget nodes with forgotten vertex $u \in X_\ell$, we maximize over all feasible choices for o^ℓ_u and m^ℓ_e with $e \in U$ and consequently $|\mathcal{L}^\ell(x^t)| \leq 2\, W$. Hence by enumerating over all possibilities, we can compute the weight of x^t in $\mathcal{O}(W)$ time. Considering join nodes, we bound the number of elements in $\mathcal{L}^{\ell,u}(x^t)$. The choices for m^ℓ and m^u are fixed by m^t. Splitting budgets, for every $i \in \{1, \ldots, k\}$ we have $b^t_i + 1$ ways to integrally split b^t_i between b^ℓ_i and b^u_i. Thus, in total there are $\mathcal{O}(\prod_{i=1}^{k} B_i)$ ways to split the budget between ℓ and u. Concerning the possible choices of o^ℓ and o^u, we have maximum freedom for $o^t_v = 1$, i.e., either o^ℓ_v or o^u_v takes value 1. Hence, the number of feasible combinations of o^ℓ and o^u is $\mathcal{O}(2^W)$. Consequently, $|\mathcal{L}^{\ell,u}(x^t)|$ is in $\mathcal{O}(2^W \prod_{i=1}^{k} B_i)$. Evaluating elements in $\mathcal{L}^{\ell,u}(x^t)$ can be done in $\mathcal{O}(W^2)$ time and we can thus compute the weight of x^t in $\mathcal{O}(W^2\, 2^W \prod_{i=1}^{k} B_i)$ time.

Combining all cases, we can compute the weight of labels in $\mathcal{O}(W^2\, 2^W \prod_{i=1}^{k} B_i)$ time. The resulting total running time of our algorithm is therefore in $\mathcal{O}(W^2\, 2^{W^2+W}\, |V| \prod_{i=1}^{k} B_i^2)$. Last but not least, we recall that a nice tree decomposition of G can be computed in linear time (Bodlaender, 1996; Kloks, 1994). □

We have just proven that Algorithm 6 solves the mBM on graphs with bounded treewidth in pseudo-polynomial time for a fixed number of budget constraints. If we additionally parameterize the mBM by the maximum budget $\bar{B} \coloneqq \max_{1 \leq i \leq k} B_i$, we get the following.

Corollary 14.21. *The mBM on graphs with bounded treewidth $tw(G) < W$ parametrized by the number of budget constraints k, the width bound W, and the maximum budget $\bar{B}$ is fixed-parameter tractable ($\mathcal{FPT}$).*

For multi-graphs, the running time analysis of Theorem 14.20 does not hold as we lose our bound $|E[X_t]| \leq W^2 - W$ on the number of edges induced by a bag $X_t \in \mathcal{X}$. Still, Algorithm 6 is out-of-the-box applicable to multi-graphs of bounded treewidth and as we are solely interested in the valid labels, we can strengthen the bound on the number of label we need to consider. For every matching $M \subseteq E(G_t)$, we have that $|M \cap E[X_t]| \leq \lfloor \frac{W}{2} \rfloor$ and thus are at most

$$\binom{|E|}{\lfloor \frac{W}{2} \rfloor} \leq \frac{|E|^{\lfloor \frac{W}{2} \rfloor}}{\lfloor \frac{W}{2} \rfloor !} \leq |E|^{\lfloor \frac{W}{2} \rfloor}.$$

possible valid choices for m^t. As a result, we can bound the number of valid labels for each tree node $t \in V(T)$ by $\mathcal{O}(|V|\, |E|^{\lfloor \frac{W}{2} \rfloor}\, 2^W \prod_{i=1}^{k} B_i)$ labels. An analogous argumentation as in Theorem 14.20 paired with this tighter bound on the number of valid labels yields the following corollary.

Corollary 14.22. *Let $G = (V, E)$ be a multi-graph of bounded treewidth $tw(G) < W$. Then the mBM with fixed k on G can be solved in $\mathcal{O}(2^{2W}\,|V|\,|E|^{\lfloor \frac{W+2}{2} \rfloor}\,\prod_{i=1}^{k} B_i^2)$ time.*

We remark that trees are simple and have treewidth one. Therefore, Theorem 14.20 translates to Theorem 14.10 on trees. Series-parallel graphs are generally multi-graphs and have a treewidth of at most two. Unfortunately, Corollary 14.22 does not directly translate to Theorem 14.6 on SP-graphs as it yields a worse running time. However, for simple SP-graphs Korenblit and Levit (2011) have shown that $|E(G)| \leq 2|V(G)| - 3$ and thus Theorem 14.20 translates to Theorem 14.6 on simple SP-graphs.

15 Minimum Color-Degree Perfect b-Matching Problems

This chapter studies the minimum color-degree perfect b-matching problem (Col-BM) which is motivated by the staff assignment for MMUs. However, to stress that there are alternative non-bipartite application of the Col-BM, we discuss an additional use case that is illustrated in Figure 15.1. Assume that an airline aims to establish new flight connections using different types of aircraft. The appropriate type of aircraft is given for every connection of interest and the number of operable connections at each airport is dictated by the takeoff and landing slots owned by the airline. As unused slots have to be returned permanently by policy so that they can be reassigned to other airlines (International Air Transport Association, 2019), all available slots at all airports have to be utilized. Operating different types of aircraft at the same airport decreases flexibility in crew scheduling while it increases the cost for the storage of spare-part. Therefore, the maximum number of different types of aircraft operated at any airport should be minimized.

In the setting above, the selection of appropriate flight connections corresponds to a perfect b-matching problem, which consists in finding an edge subset of a graph such that the vertices in the resulting subgraph have certain prespecified degrees. However, a classical b-matching neglects the diversity induced by the different types of aircraft. We can model the different types of aircraft by adding colors to the edges of the underlying graph. This leads to the Col-BM, a b-matching extension on an edge-colored graph with the objective of minimizing the maximum number of differently colored edges incident to the same node.

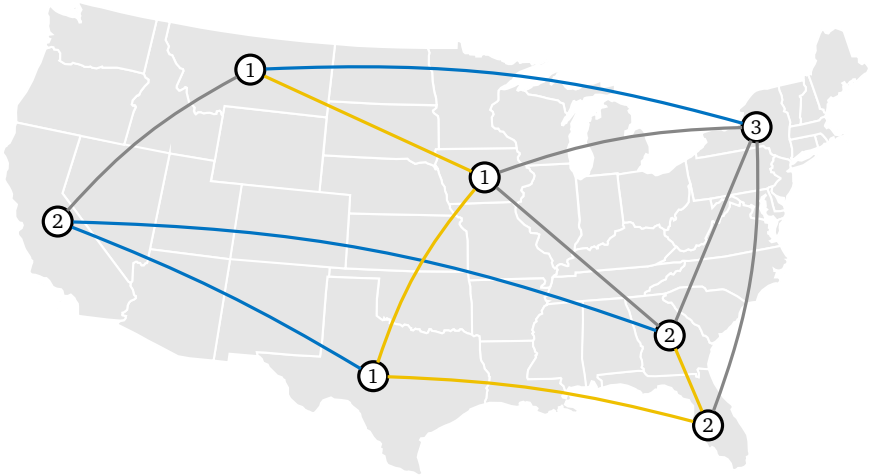


Fig. 15.1.: Exemplary Col-BM instance for the establishment of flight connections.

Before we provide a formal definition of the Col-BM, let us introduce some notation that will be used throughout this chapter. Let $G = (V, E)$ be an undirected graph with an edge coloring $E_1 \dot\cup \ldots \dot\cup E_q = E$ and $\bar{c} : E \to \{1, \ldots, q\}$ be the corresponding color function with $\bar{c}(e) = j$ if and only if $e \in E_j$. Further, let $\mathrm{col}_M(v)$ for $v \in V$ and $M \subseteq E$ denote the set of colors in $\delta_M(v)$, i.e.,

$$\mathrm{col}_M(v) \coloneqq \bar{c}(\delta_M(v)) = \{j \in \{1, \ldots, q\} \ : \ \delta_M(v) \cap E_j \neq \emptyset\}.$$

We call the number of different colors of edges in $M \subseteq E$ that are incident to $v \in V$, i.e., $|\mathrm{col}_M(v)|$, the (M-)*color degree* of v; similar to Fujita and Magnant (2011). For an edge subset $M \subseteq E$, the *color degree* of M, $f_G^{\max}(M)$, is defined as the maximum M-color degree of nodes in G, i.e., $f_G^{\max}(M) \coloneqq \max_{v \in V} |\mathrm{col}_M(v)|$. We can now formalize the Col-BM as follows.

Definition 15.1 (Col-BM)**.** Given an undirected graph $G = (V, E)$ with an edge coloring $E_1 \dot\cup \ldots \dot\cup E_q = E$, and a mapping $b \colon V \to \mathbb{N}$, the *minimum color-degree perfect b-matching problem* (Col-BM) asks for a perfect b-matching $M \subseteq E$ of minimum color degree $f_G^{\max}(M)$.

This chapter is organized as follows. In Section 15.1, we prove that the Col-BM is $\mathcal{NP}$-hard in general. However, in Section 15.2, we identify a class of two-colored complete bipartite graphs for which the Col-BM is solvable in polynomial time. Furthermore, we provide dynamic programs for the Col-BM on series-parallel graphs (Section 15.3) and on simple graphs with bounded treewidth (Section 15.4) that run in polynomial time if the number of colors is fixed.

15.1 Complexity

Concerning the complexity of the Col-BM, we remark that if $b(v) = 1$ for all $v \in V$, the Col-BM reduces to a polynomial-time solvable perfect matching problem. In the following, we show that in general the decision version of the Col-BM is strongly $\mathcal{NP}$-complete, even when restricted to $b(v) \in \{1, 2\}$ for all $v \in V$ and $q = 2$ colors.

Theorem 15.2. *The decision version of the Col-BM on two-colored bipartite graphs $G = (V_A \cup V_B, E)$ with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$ is strongly $\mathcal{NP}$-complete.*

Proof. We reduce (3,B2)-SAT to the decision version of the Col-BM. The problem (3,B2)-SAT is a strongly $\mathcal{NP}$-complete special case of 3-SAT where every literal occurs exactly twice in the formula (Berman et al., 2003). Let $\mathcal{I}$ be a (3,B2)-SAT instance with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_m$. We construct a corresponding Col-BM instance

$$\widetilde{\mathcal{I}} \coloneqq (G = ((U \cup W) \cup (V \cup R),\ E = E_1 \cup E_2),\ b),$$

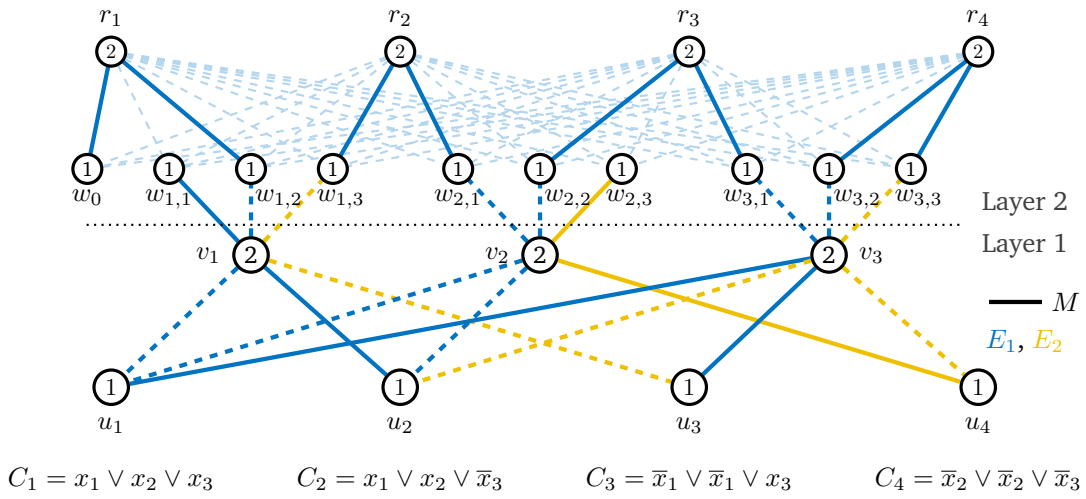


Fig. 15.2.: Construction of a perfect b-matching in the Col-BM instance $\widetilde{\mathcal{I}}$.

where G is composed of two layers; see Figure 15.2. Layer 1 models the correspondence between a perfect b-matching with color degree one and a satisfying truth assignment for an instance of (3,B2)-SAT. Layer 2 is an auxiliary, complete bipartite graph ensuring the existence of a perfect b-matching. In the following, we refer to edges in E_1 as blue edges and to edges in E_2 as yellow edges.

Layer 1 contains two sets of nodes $V \coloneqq \{v_1, \dots, v_n\}$ and $U \coloneqq \{u_1, \dots, u_m\}$, representing the variables and clauses of $\mathcal{I}$, respectively. We connect V and U via the following edges: blue edges $\{v_i, u_j\}$ for all positive literals $x_i \in C_j$ and yellow edges $\{v_i, u_j\}$ for all negative literals $\overline{x}_i \in C_j$. Finally, we set $b(v) = 2$ for all $v \in V$ and $b(u) = 1$ for all $u \in U$. As a result, Layer 1 is bipartite by construction and $\sum_{v \in V} b(v) > \sum_{u \in U} b(u)$ as for $\mathcal{I}$ holds $3m = 4n$.

Layer 2 contains two sets of nodes $W \coloneqq \{w_{i,k} : i \in \{1, \dots, n\},\ k \in \{1,2,3\}\} \cup W'$ and $R \coloneqq \{r_i : i \in \{1, \dots, \lceil \frac{7}{6}n \rceil\}\}$ that ensure the existence of a perfect b-matching in G. Note that $\frac{7}{3}n$ is integer as 3 divides n. If $\frac{7}{3}n$ is even, we define $W' \coloneqq \emptyset$ and otherwise $W' \coloneqq \{w_0\}$. We connect W with V and R via the following edges: a yellow-colored edge $\{v_i, w_{i,3}\}$ and blue-colored edges $\{v_i, w_{i,1}\}$, $\{v_i, w_{i,2}\}$ for each $i \in \{1, \dots, n\}$, as well as blue-colored edges $\{r, w\}$ for all $r \in R$ and $w \in W$. Finally, we set $b(w) = 1$ for all $w \in W$ and $b(r) = 2$ for all $r \in R$. As a result, G is bipartite by construction with node partitions $V \cup R$ and $U \cup W$, b-values $b(x) = 2$ for $x \in V \cup R$ and $b(y) = 1$ for $y \in U \cup W$, and

$$\sum_{v \in V} b(v) + \sum_{r \in R} b(r) = \sum_{u \in U} b(u) + \sum_{w \in W} b(w).$$

The Col-BM instance $\widetilde{\mathcal{I}}$ can be constructed in polynomial time. Hence, it remains to be shown that $\mathcal{I}$ is a Yes-instance if and only if $\widetilde{\mathcal{I}}$ has a perfect b-matching M with color degree $f_G^{\max}(M) = 1$.

Let $M \subseteq E$ be a perfect b-matching in G with $f_G^{\max}(M) = 1$. Then $|\mathrm{col}_M(v_i)| = 1$ for all $i \in \{1, \dots, n\}$ and we set $x_i =$ True if both edges in $\delta_M(v_i)$ are blue and $x_i =$ False

if both edges are yellow. It remains to be shown that x is a satisfying assignment for $\mathcal{I}$. By construction, for all $j \in \{1, \ldots, m\}$ there exists exactly one $i \in \{1, \ldots, n\}$ such that $\delta_M(u_j) = \{\{v_i, u_j\}\}$. If $\{v_i, u_j\}$ is blue, then $x_i \in C_j$ by construction. Hence, our choice $x_i =$ True verifies clause C_j. Analogously, if $\{v_i, u_j\}$ is yellow, then $\overline{x}_i \in C_j$. Hence, our choice $x_i =$ False verifies clause C_j. Consequently, x is a satisfying assignment for $\mathcal{I}$.

Conversely, let x be a satisfying truth assignment for $\mathcal{I}$. We construct a perfect b-matching $M \subseteq E$ as follows. We choose a verifying literal x_i ($\overline{x}_i$) for each clause C_j and add the corresponding blue (yellow) edge $\{v_i, u_j\}$ to M. Thus, we select m edges in Layer 1 and $|\delta_M(u)| = b(u) = 1$ holds for all $u \in U$. As x_i and $\overline{x}_i$ cannot simultaneously be satisfied by x, $\delta_M(v)$ contains only edges of the same color for all $v \in V$. Hence, up to this point it holds that $f_G^{\max}(M) = 1$.

To conclude our reduction, it suffices to extend M to Layer 2 without increasing $f_G^{\max}(M)$. Therefore, we proceed for every $v_i \in V$ with $|\delta_M(v_i)| < b(v_i)$ as follows: if $\delta_M(v_i) \cap E_1 \neq \emptyset$, add $\{v_i, w_{i,1}\}$ to M; if $\delta_M(v_i) \cap E_2 \neq \emptyset$, add $\{v_i, w_{i,3}\}$ to M; if $\delta_M(v_i) = \emptyset$, add both $\{v_i, w_{i,1}\}$ and $\{v_i, w_{i,2}\}$ to M. Thus, $|\delta_M(v)| = 2$ and $|\mathrm{col}_M(v)| = 1$ for all $v \in V$. Finally, let M' be a perfect b-matching in $G' \coloneqq G[R \cup \{w \in W \ : \ \delta_M(w) = \emptyset\}]$, which exists as G' is a complete bipartite graph and, by construction,

$$\sum_{r \in R} b(r) = \sum_{w \in W \colon \delta_M(w) = \emptyset} b(w).$$

Consequently, $M^* \coloneqq M \cup M'$ is a perfect b-matching in G with $f_G^{\max}(M^*) = 1$.

As the decision version of the Col-BM is obviously in $\mathcal{NP}$, as we can check the feasibility and color-degree of a given b-matching in $\mathcal{O}(|V(G)| \cdot |E|)$ time, the problem's strong $\mathcal{NP}$-completeness follows. □

Theorem 15.2 states that we can solve the strongly $\mathcal{NP}$-complete (3,B2)-SAT problem by deciding whether an optimal perfect b-matching in G has color degree one or two. This directly implies the inapproximability of the Col-BM.

Corollary 15.3. *There exists no α-approximation algorithm for the Col-BM with $1 < \alpha < 2$ unless $\mathcal{P} = \mathcal{NP}$.*

Note that any b-matching in a two-colored graph has color degree at most two. Hence, every algorithm that produces a perfect b-matching is a 2-approximation algorithm for the Col-BM on two-colored graphs. Further remark that the Col-BM on a two-colored bipartite graph $G = (V_A \cup V_B, E)$ with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$ corresponds to the task of partitioning G into monocromatic paths of length 3 whose end-nodes are exclusively in V_A (spanning P_3-partition). It is known that partitioning an uncolored graph into paths of length 3 (P_3-partition) is $\mathcal{NP}$-complete on bipartite graphs of maximum degree 3 (Monnot and Toulouse, 2007). However, to the best of our knowledge, no work has been published on monochromatic P_3-partition problems in edge-colored graphs nor on spanning P_3-partition

problems in uncolored graphs. In the case that $b(v) = r$, $r \in \mathbb{N}$, for all $v \in V$, the Col-BM is closely related to the partitioning of graphs into monochromatic r-factors. A survey on partitioning problems of edge-colored graphs into monochromatic subgraphs can be found in Kano and Li (2008).

15.2 Complete Bipartite Graphs

In the previous section, we showed that the Col-BM is $\mathcal{NP}$-hard on two-colored bipartite graphs $G = (V_A \cup V_B, E = E_1 \cup E_2)$ with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$. In this section, we additionally assume G to be complete and prove that in this case the Col-BM is solvable in polynomial time by providing a constructive algorithm. For better lucidity we abbreviate the edge notation $\{v, w\}$ as vw throughout this section.

Let $G = (V_A \cup V_B, E = E_1 \cup E_2)$ be a two-colored complete bipartite graph with color function $\bar{c}$ and $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$. We assume $|V_A| = 2\,|V_B|$ to ensure that G contains a perfect b-matching. As a result, the Col-BM reduces to the question whether G contains a perfect b-matching $M \subseteq E$ with $f_G^{\max}(M) = 1$.

We utilize two characteristics of such graphs to classify those for which a perfect b-matching $M \subseteq E$ with $f_G^{\max}(M) = 1$ exists. We begin by identifying a subgraph which is sufficient for the existence of a perfect b-matching $M \subseteq E$ with $f_G^{\max}(M) = 1$.

Lemma 15.4. *Let $G = (V_A \cup V_B, E)$ be a two-colored complete bipartite graph with $b(v) = 1$ for all $v \in V_A$, $b(v) = 2$ for all $v \in V_B$. If G contains the gadget*

$$G' \coloneqq (\{b, c, d\} \cup \{r, s\}, \{br\} \cup \{bs, cr, cs, ds\})\,,$$

illustrated in Figure 15.3(a) as a subgraph, then there exists a perfect b-matching $M \subseteq E$ in G with color degree $f_G^{\max}(M) = 1$.

Proof. Let $G = (V_A \cup V_B, E)$ be a graph that contains the subgraph G'. We present an algorithm to construct a perfect b-matching $M \subseteq E$ in G with $f_G^{\max}(M) = 1$. To that end, let G' be the subgraph defined above and initialize $M = \emptyset$. For a given edge subset $M \subseteq E$, we call a node $v \in V(G)$ (M)-*unsatisfied* if $|\delta_M(v)| < b(v)$.

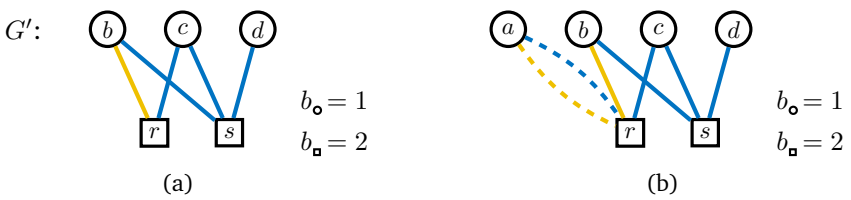


Fig. 15.3.: (a) Sufficient subgraph G'. (b) Case distinction for edge $ar \in E$.

Repeat the following two steps until all $w \in V_B \setminus V(G')$ are satisfied. First, choose an unsatisfied node $w \in V_B \setminus V(G')$ and three distinct, unsatisfied nodes $v_1, v_2, v_3 \in V_A \setminus V(G')$. Second, add two arbitrary edges $e, f \in \{v_1w, v_2w, v_3w\}$ of identical color to M, which exist as G is two-colored.

By construction $f_G^{\max}(M) = 1$, exactly one node $a \in V_A \setminus V(G')$ remains unsatisfied, and M is a perfect b-matching in $G[V(G) \setminus (\{a\} \cup V(G'))]$. Hence, it suffices to prove that there always exists a perfect b-matching M' in the induced subgraph $G'' := G[\{a\} \cup V(G')]$ with $f_{G''}^{\max}(M') = 1$, as then $M^* := M \cup M'$ is a perfect b-matching in G with $f_G^{\max}(M^*) = 1$. We distinguish two cases based on the color of the edge ar; see Figure 15.3(b):

1) If the color $\bar{c}(ar) = \bar{c}(br)$, then $M' = \{ar, br, cs, ds\}$ is a perfect b-matching in G'' with $f_{G''}^{\max}(M') = 1$.

2) If the color $\bar{c}(ar) \neq \bar{c}(br)$, then $\bar{c}(ar) = \bar{c}(cr)$ and $M' = \{ar, cr, bs, ds\}$ is a perfect b-matching in G'' with $f_{G''}^{\max}(M') = 1$.

In either case, $M^* := M \cup M'$ is a perfect b-matching in G with $f_G^{\max}(M^*) = 1$. □

As not all Col-BM instances contain the gadget G', we continue by exploiting the fact that in every perfect b-matching M in a q-colored complete bipartite graph $G = (V_A \cup V_B, E)$ with $f_G^{\max}(M) = 1$, the incident edges $\delta_M(v)$ of every node $v \in V_B$ are necessarily of the same color. We still assume $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$. As a result, for every node $v \in V_B$ only node pairs that are connected to v by edges of the same color are potential matching partners.

Definition 15.5. Let G be a q-colored graph. For $v \in V(G)$ and a color $i \in \{1, \dots, q\}$, we define the *i-colored neighborhood* of v as

$$N^i(v) := \{w \in V(G) \ : \ \bar{c}(vw) = i\}.$$

Remark that in a complete bipartite graph $G = (V_A \cup V_B, E)$, every node $v \in V_B$ induces a partition $\{N^1(v), \dots, N^q(v)\}$ of V_A. If this partition of V_A is identical for all $v \in V_B$, i.e., $\{N^1(r), \dots, N^q(r)\} = \{N^1(s), \dots, N^q(s)\}$ for all $r, s \in V_B$, we call G *stable (color) partitioned*; see Figure 15.4. We use the notion of a stable partitioning to determine whether a perfect b-Matching M in G with color degree $f_G^{\max}(M) = 1$ exists.

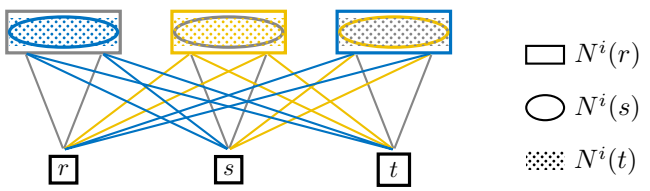


Fig. 15.4.: Stable-partitioned graph with partitions of V_A induced by the i-colored neighborhoods of nodes $r, s, t \in V_B$.

Lemma 15.6. *Let $G = (V_A \cup V_B, E)$ be a q-colored, stable-partitioned, complete bipartite graph with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$. Then there exists a perfect b-matching $M \subseteq E$ in G with $f_G^{\max}(M) = 1$ if and only if $|N^i(w)|$ is even for all colors $i \in \{1, \dots, q\}$ and all nodes $w \in V_B$.*

Proof. Let $\{P_1, \dots, P_q\}$ denote the unique partition of V_A induced by the set of i-colored neighborhoods of $r \in V_B$. If $|P_i|$ is even for all $i \in \{1, \dots, q\}$, we can construct a perfect b-matching $M \subseteq E$ with $f_G^{\max}(M) = 1$ by iteratively matching two unsatisfied nodes belonging to the same class P_i to an unsatisfied node of V_B.

Conversely, if $M \subseteq E$ is a perfect b-Matching in G with $f_G^{\max}(M) = 1$, then every P_i is canonically partitioned by M into disjoint node pairs. Thus, $|P_i|$ has to be even for all $i \in \{1, \dots, q\}$. □

We can now show that every two-colored complete bipartite graph $G = (V, E)$ with $|V| > 6$ either fulfills the conditions of Lemma 15.4 or the conditions of Lemma 15.6. This leads to a complete characterization of two-colored complete bipartite graphs with more than six nodes and will be used to derive an algorithm for the Col-BM on this graph class.

Lemma 15.7. *Let $G = (V_A \cup V_B, E)$ be a two-colored complete bipartite graph with $|V_A| = 2|V_B|$ and $|V(G)| > 6$. Then exactly one of the following is true.*

1) G contains the gadget G' defined in Lemma 15.4.

2) G is stable partitioned.

Proof. Assume 1) holds. Then 2) is violated as r and s induce different partitions of $\{b, c, d\}$.

Conversely, assume 2) is violated. Therefore, there exist $r, s \in V_B$ such that $\{N^1(r), N^2(r)\} \neq \{N^1(s), N^2(s)\}$. Hence, at least one of the following holds

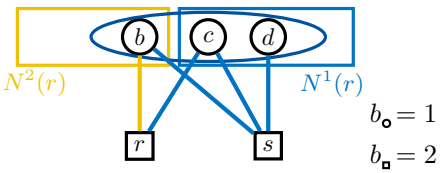
i) $N^1(s)$ intersects both $N^1(r)$ and $N^2(r)$, i.e., $N^1(r) \cap N^1(s) \neq \emptyset \ \wedge \ N^2(r) \cap N^1(s) \neq \emptyset$,

ii) $N^2(s)$ intersects both $N^1(r)$ and $N^2(r)$, i.e., $N^1(r) \cap N^2(s) \neq \emptyset \ \wedge \ N^2(r) \cap N^2(s) \neq \emptyset$.

Remark that as $|V(G)| > 6$ and $|V_A| = 2|V_B|$, it directly follows that $|V_A| \geq 6$ and $|V_B| \geq 3$.

Without loss of generality, assume that i) holds as the argumentation is analogous in the case that ii) holds. The validity of i) directly implies $|N^1(s)| \geq 2$. If $|N^1(s)| \geq 3$, then we choose $b \in N^2(r) \cap N^1(s)$, $c \in N^1(r) \cap N^1(s)$ and $d \in N^1(s) \setminus \{b, c\}$. Therefore, bs, cr, cs and ds are of color one whereas br is of color two; see Figure 15.5. Consequently,

Fig. 15.5.: Setting if i) holds and $|N^1(s)| \geq 3$.



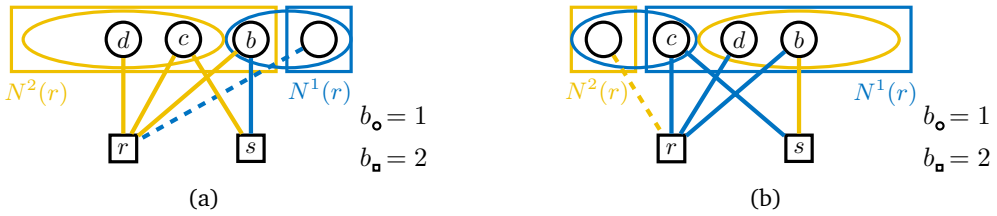


Fig. 15.6.: Setting if $|N^1(s)| = 2$ and ii) is violated and (a) $|N^1(r)| = 1$; (b) $|N^2(r)| = 1$.

$(\{b, c, d, r, s\}, \{br, bs, cr, cs, ds\})$ represents a gadget as defined in Lemma 15.4. If $|N^1(s)| = 2$ and ii) holds, then $|N^2(s)| \geq 3$ and the statement follows by symmetry.

Therefore, assume that $|N^1(s)| = 2$ and ii) is violated. Then either $|N^1(r)| = 1$ or $|N^2(r)| = 1$; see Figure 15.6. If $|N^1(r)| = 1$, then $|N^2(r)| \geq 5$ and $|N^2(r) \cap N^2(s)| \geq 4$. We choose $b \in N^2(r) \cap N^1(s)$, $c \in N^2(r) \cap N^2(s)$, and $d \in N^2(r) \setminus \{b, c\}$. Therefore, br, cr, cs and dr are of color two whereas bs is of color one. Consequently, $(\{b, c, d, r, s\}, \{br, bs, cr, cs, dr\})$ represents a gadget as defined in Lemma 15.4; see Figure 15.6(a). If $|N^2(r)| = 1$, then $|N^1(r)| \geq 5$ and $|N^1(r) \cap N^2(s)| \geq 4$, and we choose $b \in N^1(r) \cap N^2(s)$, $c \in N^1(r) \cap N^1(s)$ and $d \in N^1(r) \setminus \{b, c\}$. Therefore br, cr, cs, and dr are of color one whereas bs is of color two. Consequently, $(\{b, c, d, r, s\}, \{br, bs, cr, cs, dr\})$ represents a gadget as defined in Lemma 15.4; see Figure 15.6(b).

We conclude, if 2) is violated, then 1) holds which completes our proof. □

Remark that the condition imposed on the size of the graph G in Lemma 15.7 is tight.

Proposition 15.8. There exists a graph G with $|V(G)| = 6$ that is neither stable partitioned nor does it contain the gadget G'.

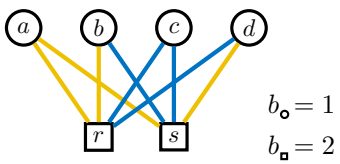
Proof. The graph

$$(\{a, b, c, d\} \cup \{r, s\}, \{ar, as, br, ds\} \cup \{bs, cr, cs, dr\})$$

is neither stable partitioned nor contains the gadget G'; see Figure 15.7. □

Lemma 15.7 implies that, on a two-colored complete bipartite graph $G = (V_A \cup V_B, E)$ with $|V(G)| > 6$, the Col-BM can be reduced to identifying the gadget G' as a subgraph, or determining that none exists; see Algorithm 7. We first check whether G is stable partitioned. If this is the case, G does not contain the gadget G' and we can determine the minimum

Fig. 15.7.: Graph G with $|V(G)| = 6$ that neither is stable partitioned nor contains the gadget G'.



Algorithm 7: The Col-BM on complete two-colored bipartite graphs

Input: Two-colored complete bipartite graph $G = (V_A \cup V_B, E = E_1 \cup E_2)$ with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$

Output: Minimum color degree of a perfect b-matching in G

```
1  choose r ∈ V_B
2  P := {N^1(r), N^2(r)}
3  for s ∈ V_B \ {r} do                          // check if G is stable partitioned
4  |  S := {N^1(s), N^2(s)}
5  |  if P ≠ S then                                    // gadget G' exists
6  |  |  return f^max := 1
7  for p ∈ P do                                  // check if all |p| are even
8  |  if |p| is odd then
9  |  |  return f^max := 2
10 return f^max := 1
```

color degree of a perfect b-matching in G by checking the cardinalities of the elements of the unique partition of V_A: if all cardinalities are even, the minimum color degree of a perfect b-matching in G is one otherwise it is two. In the other case, G contains the gadget G' and hence the minimum color degree of a perfect b-matching in G is one.

Theorem 15.9. *The Col-BM on two-colored complete bipartite graphs $G = (V_A \cup V_B, E)$ with $b(v) = 1$ for all $v \in V_A$ and $b(v) = 2$ for all $v \in V_B$ can be solved in $\mathcal{O}(|V|^2)$ time using Algorithm 7.*

Proof. The correctness of Algorithm 7 follows from Lemmas 15.4, 15.6, and 15.7. Analyzing the algorithm's running time: The partitions P and S can be computed in $\mathcal{O}(|V_A|) = \mathcal{O}(|V|)$ time. The comparison of S and P can be performed in $\mathcal{O}(|V_A|)$ time if they are represented using characteristic vectors. Thus, Algorithm 7 checks if every $w \in V_B$ induces the same partition of V_A in $\mathcal{O}(|V_A| + |V_B|\,|V_A|)$ time. The cardinalities of the two color classes and their parity can be checked in $\mathcal{O}(|V_A|)$ time. Hence, Algorithm 7 solves the Col-BM in $\mathcal{O}(|V|^2)$ time. □

Remark that if Algorithm 7 terminates in line 6 (line 10), an optimal perfect b-Matching can be determined using the construction from the proof of Lemma 15.4 (Lemma 15.6).

15.3 Series-parallel Graphs

In this section, we consider the Col-BM on series-parallel graphs. We show that, in case of a fixed number of colors, the Col-BM can be solved in polynomial time on SP-graphs by dynamic programming. Subsequently, we extend this dynamic program to solve the Col-BM

on trees. Series-parallel graphs and their decomposition trees were formally introduced in Definition 14.4.

Let $G = (V, E)$ be an SP-graph with edge coloring $E_1 \dot\cup \ldots \dot\cup E_q = E$, and $b\colon V \to \mathbb{N}$ a mapping. It is known that a decomposition tree of an SP-graph can be computed in linear time (Valdes et al., 1982). Thus, let T be a decomposition tree for G. For $t \in V(T)$, let G_t denote the subgraph of G with source σ^t and sink τ^t corresponding to t. We propose a dynamic program to solve the Col-BM on SP-graphs using the corresponding decomposition trees. First, we introduce a set of labels

$$\mathcal{L}^t \coloneqq \{\, (\alpha, F_\sigma, \beta, F_\tau) \;:\; 0 \le \alpha \le b(\sigma^t),\ 0 \le \beta \le b(\tau^t),\ F_\sigma, F_\tau \subseteq \{1, \ldots, q\}\}$$

for every tree node $t \in V(T)$. The parameters α and β define new, smaller b-values for σ^t and τ^t, whereas the *color-subsets* F_σ, F_τ define the prespecified set of colors for edges incident to σ^t and τ^t, respectively.

Before we specify the dynamic program, we introduce some additional notation. For any node $t \in V(T)$ and a label $x = (\alpha, F_\sigma, \beta, F_\tau) \in \mathcal{L}^t$, we call an edge subset $M \subseteq E(G_t)$ a (t, x)*-restricted matching* if $|\delta_M(\sigma^t)| = \alpha$, $|\delta_M(\tau^t)| = \beta$, $\mathrm{col}_M(\sigma^t) = F_\sigma$, $\mathrm{col}_M(\tau^t) = F_\tau$ and $|\delta_M(v)| = b(v)$ for all $v \in V(G_t) \setminus \{\sigma^t, \tau^t\}$. Consequently, we define the (t, x)-restricted Col-BM as

$$\min_{M \subseteq E(G_t)} \left\{ f_{G_t}^{\max}(M) \;:\; M \text{ is } (t,x)\text{-restricted matching in } G_t \right\}.$$

For a node $t \in V(T)$ and a label $x \in \mathcal{L}^t$, we call the optimal solution value of the (t, x)-restricted Col-BM the cost $c^t(x)$ of x at t. Thus, for all perfect b-matchings M^* in G with minimum color degree it holds that

$$f_G^{\max}(M^*) \;=\; \min_{F_\sigma, F_\tau \subseteq \{1,\ldots,q\}} c^r\left((b(\sigma^r), F_\sigma, b(\tau^r), F_\tau)\right),$$

for the root $r \in V(T)$ of the decomposition tree.

Our dynamic program solving the Col-BM on SP-graphs exploits the structure of decomposition trees and recursively computes label costs bottom-up. To that end, we consider the three types of nodes in the decomposition tree T of G starting with the initialization in leaves.

Lemma 15.10. *Let $t \in V(T)$ be a leaf in T, and let $e \in E$ denote the single edge in the corresponding graph G_t. Then $c^t((0, \emptyset, 0, \emptyset)) = 0$, $c^t((1, \{\bar c(e)\}, 1, \{\bar c(e)\})) = 1$, and $c^t(x) = \infty$ for all other labels $x \in \mathcal{L}^t$.*

Proof. If $t \in V(T)$ is a leaf in T, the corresponding graph G_t consists of exactly one edge by the definition of decomposition trees, i.e., $E(G_t) = \{e\}$. Therefore, there exists exactly one $(t, (0, \emptyset, 0, \emptyset))$-restricted matching: $M_0 = \emptyset$. Hence, $c^t((0, \emptyset, 0, \emptyset)) = f_{G_t}^{\max}(M_0) = 0$. Moreover, there also exists exactly one $(t, (1, \{\bar c(e)\}, 1, \{\bar c(e)\}))$-restricted matching: $M_1 = \{e\}$. Hence, $c^t((1, \{\bar c(e)\}, 1, \{\bar c(e)\})) = f_{G_t}^{\max}(M_1) = 1$. For all other labels $x \in \mathcal{L}^t$, the (t, x)-restricted Col-BM is infeasible and $c^t(x) = \infty$. □

For the two remaining types of tree nodes, we can derive the cost of labels recursively from the label costs of their child nodes. We begin by considering S-nodes, which correspond to the series composition of the graphs associated with its child nodes. As a result of this interrelation, every restricted matching at an S-nodes can be decomposed into two restricted matchings at its child nodes. By minimizing over all feasible combinations of restricted matchings at the child nodes, we get the following result.

Lemma 15.11. *Let $t \in V(T)$ be an S-node in T with child nodes ℓ and u. Then the cost of $x^t = (\alpha^t, F_\sigma^t, \beta^t, F_\tau^t) \in \mathcal{L}^t$ at t can be computed as*

$$c^t(x^t) = \min_{\substack{0 \le k \le b(\tau^\ell),\\ F_\tau^\ell, F_\sigma^u \subseteq \{1,\dots,q\}}} \max\left\{ c^\ell\left(\left(\alpha^t, F_\sigma^t, k, F_\tau^\ell\right)\right), c^u\left(\left(b(\tau^\ell) - k, F_\sigma^u, \beta^t, F_\tau^t\right)\right), \left|F_\tau^\ell \cup F_\sigma^u\right| \right\}.$$

Proof. If $t \in V(T)$ is an S-node with child nodes ℓ and u, by definition $\sigma^t = \sigma^\ell$, $\tau^t = \tau^u$, and $\tau^\ell = \sigma^u =: y$. Let $x^t = (\alpha^t, F_\sigma^t, \beta^t, F_\tau^t) \in \mathcal{L}^t$ and $M^t \subseteq E(G_t)$ be an optimal solution to the (t, x^t)-restricted Col-BM, i.e., $c^t(x^t) = f_{G_t}^{\max}(M^t)$. By defining $M^\ell := M^t \cap E(G_\ell)$ and $M^u := M^t \cap E(G_u)$, it follows that

$$f_{G_t}^{\max}(M^t) = \max\left\{ f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ |\mathrm{col}_{M^\ell}(y) \cup \mathrm{col}_{M^u}(y)| \right\}. \tag{15.1}$$

Furthermore, for $\bar{k} := |\delta_{M^\ell}(y)|$, $\bar{F}_\tau^\ell := \mathrm{col}_{M^\ell}(y)$, and $\bar{F}_\sigma^u := \mathrm{col}_{M^u}(y)$ it holds that M^ℓ is an $\left(\ell, \left(\alpha^t, F_\sigma^t, \bar{k}, \bar{F}_\tau^\ell\right)\right)$-restricted matching in G_ℓ while M^u is a $\left(u, \left(b(y) - \bar{k}, \bar{F}_\sigma^u, \beta^t, F_\tau^t\right)\right)$-restricted matching in G_u. Thus by definition, we have that $f_{G_\ell}^{\max}(M^\ell) \ge c^\ell\left(\left(\alpha^t, F_\sigma^t, \bar{k}, \bar{F}_\tau^\ell\right)\right)$ and $f_{G_u}^{\max}(M^u) \ge c^u\left(\left(b(y) - \bar{k}, \bar{F}_\sigma^u, \beta^t, F_\tau^t\right)\right)$ which yields in combination with (15.1) that

$$\begin{aligned} c^t(x^t) = f_{G_t}^{\max}(M^t) &= \max\left\{ f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ |\mathrm{col}_{M^\ell}(y) \cup \mathrm{col}_{M^u}(y)| \right\} \\ &\ge \max\left\{ c^\ell\left(\left(\alpha^t, F_\sigma^t, \bar{k}, \bar{F}_\tau^\ell\right)\right),\ c^u\left(\left(b(y) - \bar{k}, \bar{F}_\sigma^u, \beta^t, F_\tau^t\right)\right),\ \left|\bar{F}_\tau^\ell \cup \bar{F}_\sigma^u\right| \right\} \\ &\ge \min_{\substack{0 \le k \le b(y),\\ F_\tau^\ell, F_\sigma^u \subseteq \{1,\dots,q\}}} \max\left\{ c^\ell\left(\left(\alpha^t, F_\sigma^t, k, F_\tau^\ell\right)\right), c^u\left(\left(b(y) - k, F_\sigma^u, \beta^t, F_\tau^t\right)\right), \left|F_\tau^\ell \cup F_\sigma^u\right| \right\}. \end{aligned}$$

Conversely, let

$$k^*, F_\tau^{\ell *}, F_\sigma^{u*} = \operatorname*{arg\,min}_{\substack{0 \le k \le b(y),\\ F_\tau^\ell, F_\sigma^u \subseteq \{1,\dots,q\}}} \max\left\{ c^\ell\left(\left(\alpha^t, F_\sigma^t, k, F_\tau^\ell\right)\right), c^u\left(\left(b(y) - k, F_\sigma^u, \beta^t, F_\tau^t\right)\right), \left|F_\tau^\ell \cup F_\sigma^u\right| \right\}.$$

Moreover, let $M^\ell \subseteq E(G_\ell)$ be an optimal solution to the $\left(\ell, \left(\alpha^t, F_\sigma^t, k^*, F_\tau^{\ell *}\right)\right)$-restricted Col-BM on G_ℓ and $M^u \subseteq E(G_u)$ be an optimal solution to the $\left(u, (b(y) - k^*, F_\sigma^{u*}, \beta^t, F_\tau^t)\right)$-

restricted Col-BM on G_u. We define the matching $M^t \coloneqq M^\ell \cup M^u$ in G_t. By construction, M^t is (t, x^t)-restricted, $\mathrm{col}_{M^\ell}(\tau) = F_\tau^{\ell *}$, and $\mathrm{col}_{M^u}(\sigma) = F_\sigma^{u*}$. Thus,

$$\begin{aligned} c^t(x^t) &\leq f_{G_t}^{\max}(M^t) = \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|\mathrm{col}_{M^\ell}(\tau) \cup \mathrm{col}_{M^u}(\sigma)\right|\right\} \\ &= \max\left\{c^\ell\left(\left(\alpha^t, F_\sigma^t, k^*, F_\tau^{\ell *}\right)\right),\ c^u\left(\left(b(y) - k^*, F_\sigma^{u*}, \beta^t, F_\tau^t\right)\right),\ \left|F_\tau^{\ell *} \cup F_\sigma^{u*}\right|\right\} \\ &= \min_{\substack{0 \leq k \leq b(y),\\ F_\tau^\ell, F_\sigma^u \subseteq \{1,\dots,q\}}} \max\left\{c^\ell\left(\left(\alpha^t, F_\sigma^t, k, F_\tau^\ell\right)\right), c^u\left(\left(b(y) - k, F_\sigma^u, \beta^t, F_\tau^t\right)\right), \left|F_\tau^\ell \cup F_\sigma^u\right|\right\}. \end{aligned}$$

□

To conclude the computation of label costs, we consider P-nodes. Recall, that P-nodes correspond to the parallel composition of the graphs associated with its child nodes. Thus, we can similarly compute the cost of labels by minimizing over all feasible combinations of restricted matchings at the child nodes.

Lemma 15.12. *Let $t \in V(T)$ be a P-node in T with child nodes ℓ and u. Then the cost of $x^t = (\alpha^t, F_\sigma^t, \beta^t, F_\tau^t) \in \mathcal{L}^t$ at t can be computed as*

$$c^t(x^t) = \min_{\substack{0 \leq k \leq \alpha^t,\ F_\sigma^\ell \cup F_\sigma^u = F_\sigma^t \\ 0 \leq m \leq \beta^t,\ F_\tau^\ell \cup F_\tau^u = F_\tau^t}} \max\left\{c^\ell\left(\left(k, F_\sigma^\ell, m, F_\tau^\ell\right)\right), c^u\left(\left(\alpha^t - k, F_\sigma^u, \beta^t - m, F_\tau^u\right)\right), \left|F_\sigma^t\right|, \left|F_\tau^t\right|\right\}.$$

Proof. If $t \in V(T)$ is a P-node with child nodes ℓ and u, by definition $\sigma^\ell = \sigma^u = \sigma^t$ and $\tau^\ell = \tau^u = \tau^t$. Let $x^t = (\alpha^t, F_\sigma^t, \beta^t, F_\tau^t) \in \mathcal{L}^t$ and $M^t \subseteq E(G_t)$ be an optimal solution to the (t, x^t)-restricted Col-BM, i.e., $c^t(x^t) = f_{G_t}^{\max}(M^t)$. By defining $M^\ell \coloneqq M^t \cap E(G_\ell)$ and $M^u \coloneqq M^t \cap E(G_u)$, it follows that

$$\begin{aligned} f_{G_t}^{\max}(M^t) &= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|\mathrm{col}_{M^\ell}(\sigma) \cup \mathrm{col}_{M^u}(\sigma)\right|,\ \left|\mathrm{col}_{M^\ell}(\tau) \cup \mathrm{col}_{M^u}(\tau)\right|\right\} \\ &= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|F_\sigma^t\right|,\ \left|F_\tau^t\right|\right\}. \end{aligned} \tag{15.2}$$

For the choice of $\bar{k} \coloneqq |\delta_{M^\ell}(\sigma)|$, $\bar{m} \coloneqq |\delta_{M^\ell}(\tau)|$, $\bar{F}_\sigma^\ell \coloneqq \mathrm{col}_{M^\ell}(\sigma)$, and $\bar{F}_\tau^\ell \coloneqq \mathrm{col}_{M^\ell}(\tau)$, the matching M^ℓ is $\left(\ell, \left(\bar{k}, \bar{F}_\sigma^\ell, \bar{m}, \bar{F}_\tau^\ell\right)\right)$-restricted by construction. Moreover, for $\bar{F}_\sigma^u \coloneqq \mathrm{col}_{M^u}(\sigma)$ and $\bar{F}_\tau^u \coloneqq \mathrm{col}_{M^u}(\tau)$ the matching M^u is $\left(u, \left(\alpha^t - \bar{k}, \bar{F}_\sigma^u, \beta^t - \bar{m}, \bar{F}_\tau^u\right)\right)$-restricted. Thus by definition, $f_{G_\ell}^{\max}(M^\ell) \geq c^\ell\left(\left(\bar{k}, \bar{F}_\sigma^\ell, \bar{m}, \bar{F}_\tau^\ell\right)\right)$ and $f_{G_u}^{\max}(M^u) \geq c^u\left(\left(\alpha^t - \bar{k}, \bar{F}_\sigma^u, \beta^t - \bar{m}, \bar{F}_\tau^u\right)\right)$ which yields in combination with (15.2) that

$$\begin{aligned} c^t(x^t) = f_{G_t}^{\max}(M^t) &= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|F_\sigma^t\right|,\ \left|F_\tau^t\right|\right\} \\ &\geq \max\left\{c^\ell\left(\left(\bar{k}, \bar{F}_\sigma^\ell, \bar{m}, \bar{F}_\tau^\ell\right)\right),\ c^u\left(\left(\alpha^t - \bar{k}, \bar{F}_\sigma^u, \beta^t - \bar{m}, \bar{F}_\tau^u\right)\right),\ \left|F_\sigma^t\right|,\ \left|F_\tau^t\right|\right\} \\ &\geq \min_{\substack{0 \leq k \leq \alpha^t,\ F_\sigma^\ell \cup F_\sigma^u = F_\sigma^t \\ 0 \leq m \leq \beta^t,\ F_\tau^\ell \cup F_\tau^u = F_\tau^t}} \max\left\{c^\ell\left(\left(k, F_\sigma^\ell, m, F_\tau^\ell\right)\right), c^u\left(\left(\alpha^t - k, F_\sigma^u, \beta^t - m, F_\tau^u\right)\right), \left|F_\sigma^t\right|, \left|F_\tau^t\right|\right\}. \end{aligned}$$

Conversely, let $k^*, F_\sigma^{\ell*}, F_\sigma^{u*}, m^*, F_\tau^{\ell*}, F_\tau^{u*}$ be an optimal solution to

$$\underset{\substack{0\leq k\leq \alpha^t,\ F_\sigma^\ell \cup F_\sigma^u = F_\sigma^t \\ 0\leq m\leq \beta^t,\ F_\tau^\ell \cup F_\tau^u = F_\tau^t}}{\arg\min} \max\left\{c^\ell\left(\left(k, F_\sigma^\ell, m, F_\tau^\ell\right)\right), c^u\left(\left(\alpha^t - k, F_\sigma^u, \beta^t - m, F_\tau^u\right)\right), \left|F_\sigma^t\right|, \left|F_\tau^t\right|\right\}.$$

Moreover, let $M^\ell \subseteq E(G_\ell)$ be an optimal solution to the $\left(\ell, \left(k^*, F_\sigma^{\ell*}, m^*, F_\tau^{\ell*}\right)\right)$-restricted Col-BM on G_ℓ and $M^u \subseteq E(G_u)$ be an optimal solution to the $\left(u, (\alpha^t - k^*, F_\sigma^{u*}, \beta^t - m^*, F_\tau^{u*})\right)$-restricted Col-BM on G_u. We define the matching $M^t \coloneqq M^\ell \cup M^u$ in G_t. By construction, M^t is (t, x^t)-restricted, $F_\sigma^{\ell*} \cup F_\sigma^{u*} = F_\sigma^t$, and $F_\tau^{\ell*} \cup F_\tau^{u*} = F_\tau^t$. Thus,

$$\begin{aligned}
c^t(x^t) &\leq f_{G_t}^{\max}(M^t) \\
&= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|\mathrm{col}_{M^\ell}(\sigma) \cup \mathrm{col}_{M^u}(\sigma)\right|,\ \left|\mathrm{col}_{M^\ell}(\tau) \cup \mathrm{col}_{M^u}(\tau)\right|\right\} \\
&= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|F_\sigma^{\ell*} \cup F_\sigma^{u*}\right|,\ \left|F_\tau^{\ell*} \cup F_\tau^{u*}\right|\right\} \\
&= \max\left\{f_{G_\ell}^{\max}(M^\ell),\ f_{G_u}^{\max}(M^u),\ \left|F_\sigma^t\right|,\ \left|F_\tau^t\right|\right\} \\
&= \max\left\{c^\ell\left(\left(k^*, F_\sigma^{\ell*}, m^*, F_\tau^{\ell*}\right)\right),\ c^u\left(\left(\alpha^t - k^*, F_\sigma^{u*}, \beta^t - m^*, F_\tau^{u*}\right)\right),\ \left|F_\sigma^t\right|, \left|F_\tau^t\right|\right\} \\
&= \min_{\substack{0\leq k\leq \alpha^t,\ F_\sigma^\ell \cup F_\sigma^u = F_\sigma^t \\ 0\leq m\leq \beta^t,\ F_\tau^\ell \cup F_\tau^u = F_\tau^t}} \max\left\{c^\ell\left(\left(k, F_\sigma^\ell, m, F_\tau^\ell\right)\right), c^u\left(\left(\alpha^t - k, F_\sigma^u, \beta^t - m, F_\tau^u\right)\right), \left|F_\sigma^t\right|, \left|F_\tau^t\right|\right\}.
\end{aligned}$$

□

A perfect b-matching $M^* \subseteq E$ in G of minimum color degree can be obtained by backtracking the chosen minima in the steps of the dynamic program.

Next, we consider the running time of our dynamic program. For better lucidity, let $B \coloneqq \max_{v\in V} b(v)$ denote the maximum b-value.

Theorem 15.13. *The Col-BM parameterized by the number of colors q on SP-graphs is $\mathcal{FPT}$ and can be solved in $\mathcal{O}(|E| \cdot 36^q \cdot B^4)$ time.*

Proof. The correctness of the dynamic program follows from Lemmas 15.10, 15.11, and 15.12. Regarding its running time, observe that the costs of $\mathcal{O}\left(B^2 \cdot 4^q\right)$ labels need to be computed for each node $t \in V(T)$. The computational complexity of computing the costs of labels is dominated by the computation of label costs for P-nodes. For P-nodes, we have to minimize over $\mathcal{O}(B)$ choices for k and m, respectively. For each color in F_σ^t, that color can be either in F_σ^ℓ, in F_σ^u, or in both which yields $\mathcal{O}(3^q)$ possibilities. The same estimation holds for F_τ^t and thus we compute the minimum of at most $\mathcal{O}\left(9^q \cdot B^2\right)$ maxima and every maximum can be calculated in $\mathcal{O}(1)$ time. As $|V(T)| = 2\,|E| - 1$ and a decomposition tree can be computed in linear time (Valdes et al., 1982), the total running time of the algorithm is in $\mathcal{O}\left(|E| \cdot 36^q \cdot B^4\right)$.

□

We note that in all Col-BM instances, $B \leq |E|$ and therefore our algorithm has polynomial running time if q is constant.

Moreover, we can extend our algorithm to solve the Col-BM on trees using the graph transformation from Chapter 14: given a Col-BM instance $\mathcal{I}$ on a tree $T = (V, E)$, we construct an auxiliary graph G^T by adding a new vertex y, connecting it to all leaves of T and setting $b(y) = 0$. By construction, G^T is series-parallel (Lemma 14.9) and contains at most $2(|V| - 1)$ edges. Furthermore, every perfect b-matching in G^T contains no edges from $\delta_{G^T}(y) = E(G^T) \setminus E$ and is therefore a perfect b-matching in T. The edges in $\delta_{G^T}(y)$ can be colored arbitrarily.

Corollary 15.14. *The Col-BM parameterized by the number of colors q on trees is $\mathcal{FPT}$ and can be solved in $\mathcal{O}(|V| \cdot 36^q \cdot B^4)$ time.*

15.4 Graphs with Bounded Treewidth

We proceed by considering the Col-BM on graphs with bounded treewidth, which is a more general graph class that includes SP-graphs and trees. Using dynamic programming, we show that the Col-BM on graphs with bounded treewidth is polynomial-time solvable for a fixed number of colors. To that end we make use of the concept of tree decompositions and the set of nice tree decompositions that we previously formalized in Definitions 14.11 and 14.12.

Our dynamic program for solving the Col-BM on graphs with bounded treewidth exploits the structure of nice tree decompositions and recursively computes label costs bottom-up. Let $G = (V, E)$ be a graph with bounded treewidth $tw(G) < W \in \mathbb{N}$, $E_1 \dot\cup \ldots \dot\cup E_q = E$ be an edge coloring of G, and $\bar{c}\colon E \to \{1, \ldots, q\}$ the corresponding color function. Further, let $(T, \mathcal{X})$ be a nice tree decomposition of G such that $tw(G, (T, \mathcal{X})) < W$. Without loss of generality, we assume that the bag X_r, corresponding to the root r of T, contains exactly one vertex. Should $(T, \mathcal{X})$ violate this assumption, we simply add a sequence of forget nodes to r and redefine T's root.

For a tree node $t \in V(T)$ we denote the set of edges of G induced by its bag X_t with $E[X_t]$ and the subgraph of G induced by the vertices in the bags of the subtree of T rooted in t with G_t. As before, for a vertex $v \in V$ and a subset of edges $M \subseteq E$, we denote the set of colors in $\delta_M(v)$ by $\mathrm{col}_M(v)$. Finally, for all mappings $f\colon A \to B$, we abbreviate $f_a \coloneqq f(a)$ for $a \in A$ for ease of notation.

We introduce labels of the form

$$x = (m, F, \beta) \in \mathcal{L}^t \coloneqq \{0, 1\}^{E[X_t]} \times \mathcal{P}(\{1, \ldots, q\})^{X_t} \times \mathbb{N}^{X_t}$$

at the tree nodes $t \in V(T)$ to define an auxiliary variant of Col-BM on the subgraph G_t to which we refer as the xCol-BM(t,x). To that end, the binary-valued mapping $m\colon E[X_t] \to$

$\{0,1\}$ prespecifies whether an edge $e \in E[X_t]$ is part of the b-matching in G_t or not. The mapping $F\colon X_t \to \mathcal{P}(\{1,\dots,q\})$ indicates for each vertex $v \in X_t$ the set of *unlocked* edge colors $F_v \subseteq \{1,\dots,q\}$. Only edges from $\delta_{G_t}(v)$ with unlocked colors may be chosen as part of a b-matching and all unlocked colors count towards the color degree of a vertex – even if they are unused. This gives rise to the definition of the *x-(M-)color degree* of $v \in V(G_t)$:

$$|\mathrm{col}_M(x,v)| = \begin{cases} |F_v| & \text{if } v \in X_t, \\ |\mathrm{col}_M(v)| & \text{else.} \end{cases}$$

Finally, the mapping $\beta\colon X_t \to \mathbb{N}$ defines the required degree of each vertex $v \in X_t$ with respect to matching edges in $E(G_t) \setminus E[X_t]$. We formalize the auxiliary problem xCol-BM(t,x) as follows:

$$\begin{aligned}
\min_{M \subseteq E(G_t)} \quad & \max_{v \in V(G_t)} |\mathrm{col}_M(x,v)| && && \text{(15.3a)}\\
\text{s.t.} \quad & |e \cap M| = m_e && \forall e \in E[X_t] && \text{(15.3b)}\\
& |\delta_M(v)| = b(v) && \forall v \in V(G_t) \setminus X_t && \text{(15.3c)}\\
& |\delta_M(v)| \leq b(v) && \forall v \in X_t && \text{(15.3d)}\\
& |\delta_M(v) \setminus E[X_t]| = \beta_v && \forall v \in X_t && \text{(15.3e)}\\
& \mathrm{col}_M(v) \subseteq F_v && \forall v \in X_t. && \text{(15.3f)}
\end{aligned}$$

Every b-matching in G_t satisfying the constraints (15.3b) – (15.3f) is called *(t,x)-feasible*. We define the cost $c^t(x)$ of label $x \in \mathcal{L}^t$ at tree node t as the optimal solution value to the xCol-BM(t,x). If the xCol-BM(t,x) is infeasible, we call the label x *invalid* and we set $c^t(x) = \infty$. All remaining labels are called *valid* and we calculate their cost recursively. To that end, we consider the four types of nodes in the nice tree decomposition $(T, \mathcal{X})$ of G starting with the initialization in leaves.

Lemma 15.15. *Let $t \in V(T)$ be a leaf with $X_t = \{v\}$ for some $v \in V$. Then the cost of a valid label $x = (m, F, \beta) \in \mathcal{L}^t$ at t can be computed as*

$$c^t(x) = |F_v|\,.$$

Proof. As t is a leaf, $E[X_t] = \emptyset$ and G_t consists of the isolated vertex $v \in X_t$. All valid labels $x \in \mathcal{L}^t$ have the form $x = (m, F, \beta)$ with $F_v \subseteq \{1,\dots,q\}$, and $\beta_v = 0$. The only (t,x)-feasible b-matching in G_t is $M \coloneqq \emptyset$ and thus, $c^t(x) = |\mathrm{col}_M(x,v)| = |F_v|$. □

For the three remaining types of tree nodes, label costs can be derived recursively from the label costs of child nodes. We begin by considering introduce nodes.

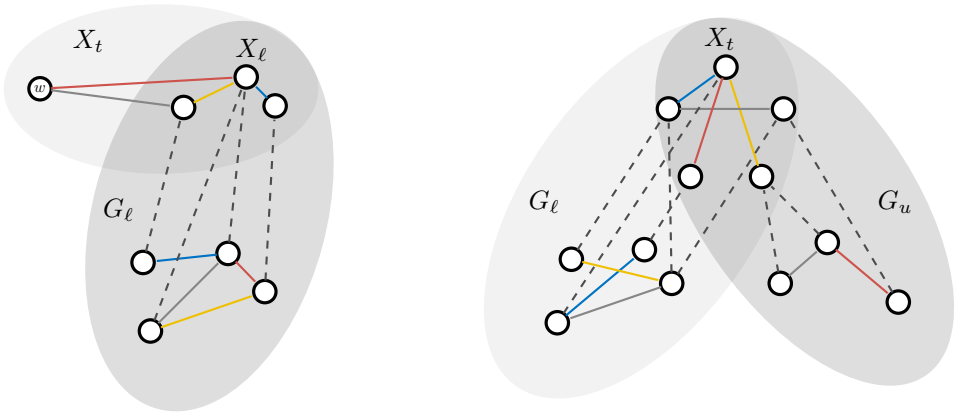


Fig. 15.8.: (a) Visualization of G_t for $t \in V(T)$ being an introduce node with child node $\ell \in V(T)$. (b) Visualization of G_t for $t \in V(T)$ being a join node with child nodes $\ell, u \in V(T)$.

Lemma 15.16. *Let $t \in V(T)$ be an introduce node with unique child node $\ell \in V(T)$, and let $w \in V$ be the introduced vertex, i.e., $X_t \setminus X_\ell = \{w\}$; see Figure 15.8(a). Given a valid label $x^t = (m^t, F^t, \beta^t) \in \mathcal{L}^t$, we define the label $x^\ell = (m^\ell, F^\ell, \beta^\ell) \in \mathcal{L}^\ell$ via $m^\ell_e := m^t_e$ for all $e \in E[X_\ell]$, $F^\ell_v := F^t_v$ for all $v \in X_\ell$, and $\beta^\ell_v := \beta^t_v$ for all $v \in X_\ell$. Then the cost of x^t at t can be computed as*

$$c^t(x^t) = \max\left\{c^\ell(x^\ell), \left|F^t_w\right|\right\}.$$

Proof. We begin by showing that $c^t(x^t) \geq \max\{c^\ell(x^\ell), |F^t_w|\}$. Let $M^t \subseteq E(G_t)$ be an optimal solution to the xCol-BM(t, x^t). For the vertex $w \in V$ introduced by node $t \in V(T)$, it holds that $\left|\mathrm{col}_{M^t}(x^t, w)\right| = \left|F^t_w\right|$ as $w \in X_t$. Hence,

$$c^t(x^t) = \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \geq \left|\mathrm{col}_{M^t}(x^t, w)\right| = \left|F^t_w\right|. \tag{15.4}$$

Next, let $U := \delta_{G_t}(w) \subseteq E[X_t]$ be the set of edges introduced by $t \in V(T)$ and $M^\ell := M^t \setminus \{U\}$. We show that M^ℓ is an (ℓ, x^ℓ)-feasible b-matching in order to bound $c^\ell(x^\ell)$ from above.

By construction of M^ℓ, $e \cap M^\ell = e \cap M^t = m^t_e = m^\ell_e$ holds for all $e \in E[X_\ell]$ and thus equalities (15.3b) are satisfied. Concerning equalities (15.3c), it holds that $|\delta_{M^\ell}(v)| = |\delta_{M^t}(v)| = b(v)$ for $v \in V(G_\ell) \setminus X_\ell$. Finally, as $M^\ell \subseteq M^t$ and M^t is (t, x^t)-feasible, it follows that

$$\begin{aligned}
&|\delta_{M^\ell}(v)| \leq |\delta_{M^t}(v)| \leq b(v) && \forall v \in X_\ell,\\
&|\delta_{M^\ell}(v) \setminus E[X_\ell]| = |\delta_{M^t}(v) \setminus E[X_t]| = \beta^t_v = \beta^\ell_v && \forall v \in X_\ell,\\
&\mathrm{col}_{M^\ell}(v) \subseteq \mathrm{col}_{M^t}(v) \subseteq F^t_v = F^\ell_v && \forall v \in X_\ell,
\end{aligned}$$

and hence conditions (15.3d) – (15.3f) are satisfied. Therefore, M^ℓ is (ℓ, x^ℓ)-feasible.

Additionally, $\left|\mathrm{col}_{M^\ell}(x^\ell, v)\right| = |\mathrm{col}_{M^t}(x^t, v)|$ for all $v \in V(G_\ell)$ as $M^t \setminus E[X_t] = M^\ell \setminus E[X_\ell]$ and $F_v^t = F_v^\ell$ for all $v \in X_\ell$. As a result,

$$\begin{aligned} c^\ell(x^\ell) &\leq \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right| = \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \\ &\leq \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| = c^t(x^t). \end{aligned} \tag{15.5}$$

By combining inequalities (15.4) and (15.5), we obtain $c^t(x^t) \geq \max\{c^\ell(x^\ell), |F_w^t|\}$.

Conversely, we show that $c^t(x^t) \leq \max\{c^\ell(x^\ell), |F_w^t|\}$. Let $M^\ell \subseteq E(G_\ell)$ be an optimal solution to the xCol-BM(ℓ, x^ℓ). We define the b-matching $M^t \coloneqq M^\ell \cup \{e \in U \,:\, m_e^t = 1\}$ and show that M^t is (t, x^t)-feasible in order to bound $c^t(x^t)$ from above. For $e \in U$, equation (15.3b) holds by definition. For $e \in E[X_t] \setminus U = E[X_\ell]$, equation (15.3b) is satisfied as $e \cap M^t = e \cap M^\ell = m_e^\ell = m_e^t$ holds. Concerning equations (15.3c), $|\delta_{M^t}(v)| = |\delta_{M^\ell}(v)| = b(v)$ holds for all $v \in V(G_t) \setminus X_t$. For the introduced vertex $w \in X_t$, constraints (15.3e) and (15.3f) hold by the validity of x^t. For $v \in X_t \setminus \{w\} = X_\ell$, equation (15.3e) holds as

$$|\delta_{M^t}(v) \setminus E[X_t]| = |\delta_{M^\ell}(v) \setminus E[X_\ell]| = \beta_v^\ell = \beta_v^t.$$

By the validity of x^t, condition (15.3f) holds for $v \in X_t \setminus \{w\}$ as

$$\mathrm{col}_{M^t}(v) = \mathrm{col}_{M^\ell}(v) \cup \mathrm{col}_{M^t \setminus M^\ell}(v) \subseteq F_v^\ell \cup F_v^t = F_v^t.$$

Finally, equations (15.3b) and (15.3e) in combination with the validity of x^t imply that inequalities (15.3d) hold for all $v \in V(G_t)$. Therefore, M^t is (t, x^t)-feasible.

The construction of M^t implies that $\left|\mathrm{col}_{M^t}(x^t, v)\right| = \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|$ for all $v \in V(G_\ell)$. Thus,

$$\begin{aligned} c^t(x^t) &\leq \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| = \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^t}(x^t, v)\right|, \left|\mathrm{col}_{M^t}(x^t, w)\right| \right\} \\ &= \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|, \left|F_w^t\right| \right\} = \max\left\{c^\ell(x^\ell), \left|F_w^t\right| \right\}. \end{aligned}$$

We conclude $c^t(x^t) = \max\{c^\ell(x^\ell), |F_w^t|\}$ which completes our proof. □

Next, we consider the computation of label costs for forget nodes.

Lemma 15.17. *Let $t \in V(T)$ be a forget node with unique child node $\ell \in V(T)$. Let $w \in V$ be the forgotten vertex, i.e., $\{w\} = X_\ell \setminus X_t$, and denote its incident edges with respect to $G[X_\ell]$ by*

$U \coloneqq E[X_\ell] \setminus E[X_t] = \delta_{G_\ell}(w) \cap E[X_\ell]$. *Given a valid label* $x^t = (m^t, F^t, \beta^t) \in \mathcal{L}^t$*, we define the set* $\mathcal{L}^\ell(x^t) \subsetneq \mathcal{L}^\ell$ *of labels at* ℓ *via*

$$\begin{aligned}\mathcal{L}^\ell(x^t) \coloneqq \Big\{(m^\ell, F^\ell, \beta^\ell) \in \mathcal{L}^\ell : m^\ell_e = m^t_e \qquad & \forall e \in E[X_t],\\ F^\ell_v = F^t_v \qquad & \forall v \in X_t,\\ \beta^\ell_v = \beta^t_v - \textstyle\sum_{e\in\delta_U(v)} m^\ell_e \qquad & \forall v \in X_t,\\ \beta^\ell_w = b(w) - \textstyle\sum_{e\in U} m^\ell_e \qquad & \Big\}.\end{aligned}$$

Then the cost of x^t *at* t *can be computed as*

$$c^t(x^t) = \min_{x^\ell \in \mathcal{L}^\ell(x^t)} c^\ell(x^\ell).$$

Proof. We begin by showing that $c^t(x^t) \geq \min_{x^\ell\in\mathcal{L}^\ell(x^t)} c^\ell(x^\ell)$. To that end, note that $G_t = G_\ell$ and let $M^t \subseteq E(G_t)$ be an optimal solution to the xCol-BM(t,x^t). We define a label $x^\ell = (m^\ell, F^\ell, \beta^\ell) \in \mathcal{L}^\ell$ as follows:

$$m^\ell_e \coloneqq \begin{cases} m^t_e & \text{for } e \in E[X_t],\\ |e \cap M^t| & \text{for } e \in U,\end{cases}$$

$$F^\ell_v \coloneqq \begin{cases} F^t_v & \text{for } v \in X_t,\\ \mathrm{col}_{M^t}(v) & \text{for } v = w,\end{cases}$$

$$\beta^\ell_v \coloneqq \begin{cases} \beta^t_v - \sum_{e\in\delta_U(v)} m^\ell_e & \text{for } v \in X_t,\\ b(v) - \sum_{e\in U} m^\ell_e & \text{for } v = w.\end{cases}$$

By construction, $x^\ell \in \mathcal{L}^\ell(x^t)$. We show that M^t is (ℓ, x^ℓ)-feasible in order to bound $c^\ell(x^\ell)$ from above. By the definition of x^ℓ, equations (15.3b) are satisfied for all $e \in U$, whereas $|e \cap M^t| = m^t_e = m^\ell_e$ for $e \in E[X^t]$ since M^t is (t, x^t)-feasible. As $V(G_\ell) \setminus X_\ell \subseteq V(G_t) \setminus X_t$, constraints (15.3c) and (15.3d) hold by the x^t-feasibility of M^t. Concerning equalities (15.3e), for any vertex $v \in X_t$

$$\big|\delta_{M^t}(v) \setminus E[X_\ell]\big| = \big|\delta_{M^t}(v) \setminus E[X_t]\big| - \big|\delta_{M^t}(v) \cap U\big| = \beta^t_v - \sum_{e\in\delta_U(v)} m^\ell_e = \beta^\ell_v,$$

whereas for the forgotten vertex w it holds that

$$\big|\delta_{M^t}(w) \setminus E[X_\ell]\big| = \big|\delta_{M^t}(w) \setminus (M^t \cap U)\big| = b(w) - \sum_{e\in U} m^\ell_e = \beta^\ell_w.$$

Finally, as M^t is (t, x^t)-feasible, constraints (15.3f) are satisfied by our definition of x^ℓ and thus M^t is (ℓ, x^ℓ)-feasible.

For the forgotten vertex $w \in X_\ell$, $F^\ell_w = \mathrm{col}_{M^t}(w)$ by our definition of x^ℓ. Thus, we conclude that $|\mathrm{col}_{M^t}(x^t, v)| = \left|\mathrm{col}_{M^t}(x^\ell, v)\right|$ for all $v \in V(G_t)$ and it follows that

$$\begin{aligned} c^t(x^t) &= \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| = \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^t}(x^\ell, v)\right| \\ &\geq c^\ell(x^\ell) \geq \min_{\tilde{x}^\ell \in \mathcal{L}^\ell(x^t)} c^\ell(\tilde{x}^\ell). \end{aligned}$$

Conversely, for all labels $x^\ell \in \mathcal{L}^\ell(x^t)$, the xCol-BM($t$,$x^t$) is a relaxation of the xCol-BM(ℓ,x^ℓ) and therefore $c^t(x^t) \leq \min_{x^\ell \in \mathcal{L}^\ell(x^t)} c^\ell(x^\ell)$.

We conclude $c^t(x^t) = \min_{x^\ell \in \mathcal{L}^\ell(x^t)} c^\ell(x^\ell)$ which completes our proof. □

To conclude the computation of label costs, we consider join nodes.

Lemma 15.18. *Let $t \in V(T)$ be a join node with child nodes ℓ and u. Given a valid label $x^t = (m^t, F^t, \beta^t) \in \mathcal{L}^t$, we define the set $\mathcal{L}^{\ell,u}(x^t) \subsetneq \mathcal{L}^\ell \times \mathcal{L}^u$ of pairs of labels at ℓ and u via*

$$\begin{aligned} \mathcal{L}^{\ell,u}(x^t) \coloneqq \Big\{ ((m^\ell,F^\ell,\beta^\ell),(m^u,F^u,\beta^u)) \in \mathcal{L}^\ell \times \mathcal{L}^u \;:\; & m^\ell_e = m^u_e = m^t_e && \forall e \in E[X_t], \\ & F^\ell_v = F^u_v = F^t_v && \forall v \in X_t, \\ & \beta^\ell_v + \beta^u_v = \beta^t_v && \forall v \in X_t \Big\}. \end{aligned}$$

Then the cost of x^t at t can be computed as

$$c^t(x^t) = \min_{(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)} \max\left\{ c^\ell(x^\ell), c^u(x^u) \right\}.$$

Proof. Recall that for join nodes $X_t = X_\ell = X_u$ and $(V(G_\ell) \setminus X_\ell) \cap (V(G_u) \setminus X_u) = \emptyset$; see Figure 15.8(b).

We begin by showing that $c^t(x^t) \geq \min_{(x^\ell,x^u) \in \mathcal{L}^{\ell,u}(x^t)} \max\{c^\ell(x^\ell), c^u(x^u)\}$. Let $M^t \subseteq E(G_t)$ be an optimal solution to the xCol-BM(t,x^t). We define the restrictions of M^t to the subgraphs G_ℓ and G_u as $M^\ell \coloneqq M^t \cap E(G_\ell)$ and $M^u \coloneqq M^t \cap E(G_u)$, respectively. Moreover, we define labels $x^\ell \coloneqq (m^t, F^t, \beta^\ell) \in \mathcal{L}^\ell$ and $x^u \coloneqq (m^t, F^t, \beta^u) \in \mathcal{L}^u$ such that $\beta^\ell_v = |\delta_{M^\ell}(v) \setminus E[X_t]|$ and $\beta^u_v = |\delta_{M^u}(v) \setminus E[X_t]|$ for all $v \in X_t$. The x^t-feasibility of M^t implies for all vertices $v \in X_t$ that

$$\beta^\ell_v + \beta^u_v = |\delta_{M^\ell}(v) \setminus E[X_t]| + |\delta_{M^u}(v) \setminus E[X_t]| = |\delta_{M^t}(v) \setminus E[X_t]| = \beta^t_v,$$

and consequently it follows that $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$.

By construction, the b-matchings M^ℓ and M^u are feasible for the xCol-BM(ℓ,x^ℓ) and the xCol-BM(u,x^u), respectively. Moreover, as $F^\ell_v = F^u_v = F^t_v$ for all $v \in X_t$, it follows that

$|\mathrm{col}_{M^t}(x^t, v)| = \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|$ for all $v \in V(G_\ell)$ and $|\mathrm{col}_{M^t}(x^t, v)| = |\mathrm{col}_{M^u}(x^u, v)|$ for all $v \in V(G_u)$. Hence, we have that

$$\begin{aligned}
c^t(x^t) &= \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \\
&= \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^t}(x^t, v)\right|, \max_{v \in V(G_u)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \right\} \\
&= \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|, \max_{v \in V(G_u)} |\mathrm{col}_{M^u}(x^u, v)| \right\} \\
&\geq \max\left\{c^\ell(x^\ell), c^u(x^u)\right\} \\
&\geq \min_{(\tilde{x}_\ell, \tilde{x}_u) \in \mathcal{L}^{\ell,u}(x^t)} \max\left\{c^\ell(\tilde{x}_\ell), c^u(\tilde{x}_u)\right\}.
\end{aligned}$$

Conversely, we show that $c^t(x^t) \leq \min_{(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)} \max\{c^\ell(x^\ell), c^u(x^u)\}$. To that end, consider a pair of labels $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$, and let $M^\ell \in E(G_\ell)$ and $M^u \in E(G_u)$ be optimal solutions to the xCol-BM(ℓ,x^ℓ) and the xCol-BM(u,x^u), respectively. We define the b-matching $M^t \coloneqq M^\ell \cup M^u$ in G_t and show that M^t is (t, x^t)-feasible. Equations (15.3b) hold, as $|e \cap M^t| = \left|e \cap M^\ell\right| = m^\ell_e = m^t_e$ for all $e \in E[X_t]$. For all $v \in V(G_\ell) \setminus X_t$ it holds that $|\delta_{M^t}(v)| = |\delta_{M^\ell}(v)| = b(v)$ and analogously $|\delta_{M^t}(v)| = |\delta_{M^u}(v)| = b(v)$ for all $v \in V(G_u) \setminus X_t$, proving that equations (15.3c) hold. Concerning constraints (15.3e), for every $v \in X_t$

$$|\delta_{M^t}(v) \setminus E[X_t]| = |\delta_{M^\ell}(v) \setminus E[X_t]| + |\delta_{M^u}(v) \setminus E[X_t]| = \beta^\ell_v + \beta^u_v = \beta^t_v$$

which, in combination with the validity of x^t, directly implies that inequalities (15.3d) are satisfied. Finally, for all $v \in X_t$ it holds that $\mathrm{col}_{M^t}(v) = \mathrm{col}_{M^\ell}(v) \cup \mathrm{col}_{M^u}(v) \subseteq F^\ell_v \cup F^u_v = F^t_v$ proving the validity of constraints (15.3f). Therefore, M^t is (t, x^t)-feasible.

Additionally, we have $|\mathrm{col}_{M^t}(x^t, v)| = \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|$ for all $v \in V(G_\ell)$ and $|\mathrm{col}_{M^t}(x^t, v)| = |\mathrm{col}_{M^u}(x^u, v)|$ for all $v \in V(G_u)$ as $F^\ell_v = F^u_v = F^t_v$ for all $v \in X_t$. We thus conclude that for all $(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)$

$$\begin{aligned}
c^t(x^t) &\leq \max_{v \in V(G_t)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \\
&= \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^t}(x^t, v)\right|, \max_{v \in V(G_u)} \left|\mathrm{col}_{M^t}(x^t, v)\right| \right\} \\
&= \max\left\{ \max_{v \in V(G_\ell)} \left|\mathrm{col}_{M^\ell}(x^\ell, v)\right|, \max_{v \in V(G_u)} |\mathrm{col}_{M^u}(x^u, v)| \right\} \\
&= \max\left\{c^\ell(x^\ell), c^u(x^u)\right\}
\end{aligned}$$

and therefore, it holds in particular that

$$c^t(x^t) \leq \min_{(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)} \max\left\{c^\ell(x^\ell), c^u(x^u)\right\}.$$

We conclude $c^t(x^t) = \min_{(x^\ell, x^u) \in \mathcal{L}^{\ell,u}(x^t)} \max\{c^\ell(x^\ell), c^u(x^u)\}$ which was to be shown. □

Finally, we show how the optimal solution value to the Col-BM on G is obtained from the computed label costs.

Lemma 15.19. *Let r be T's root with $X_r = \{z\}$, and M^* a perfect b-matching in G of minimum color degree. We define the set $\mathcal{L}^* = \{(m, F, \beta) \in \mathcal{L}^r \ : \ \beta_z = b(z)\} \subsetneq \mathcal{L}^r$ of valid labels at r. Then*

$$f_G^{\max}(M^*) = \min_{x \in \mathcal{L}^*} c^r(x).$$

Proof. First, we show that $f_G^{\max}(M^*) \geq \min_{x \in \mathcal{L}^*} c^r(x)$. To that end, consider the label $x^r = (m^r, F^r, \beta^r) \in \mathcal{L}^*$ with $F_z^r \coloneqq \mathrm{col}_{M^*}(z)$. Then M^* is by construction (r, x^r)-feasible and thus

$$\min_{x \in \mathcal{L}^*} c^r(x) \leq c^r(x^r) \leq \max_{v \in V(G_t)} |\mathrm{col}_{M^*}(x^r, v)| = \max_{v \in V} |\mathrm{col}_{M^*}(v)| = f_G^{\max}(M^*).$$

Conversely, we show that $f_G^{\max}(M^*) \leq \min_{x \in \mathcal{L}^*} c^r(x)$. Let $x^r \in \mathcal{L}^*$ and $M^r \subseteq E(G_r) = E$ be an optimal solution to the xCol-BM(r, x^r). We note that equations (15.3c) and (15.3e) ensure that M^r is a perfect b-matching in G. Therefore it holds that

$$c^r(x^r) = \max_{v \in V(G_t)} |\mathrm{col}_{M^r}(x^r, v)| \geq \max_{v \in V(G_t)} |\mathrm{col}_{M^r}(v)| = f_G^{\max}(M^r) \geq f_G^{\max}(M^*).$$

As x^r was chosen arbitrarily from $\mathcal{L}^*$, in particular we have that

$$\min_{x \in \mathcal{L}^*} c^r(x) \geq f_G^{\max}(M^*).$$

We conclude $f_G^{\max}(M^*) = \min_{x \in \mathcal{L}^*} c^r(x)$ which completes our proof. □

A perfect b-matching M^* in G of minimum color degree can be obtained by backtracking the chosen minima in the steps of the dynamic program. We can now formulate the main result of this section. For better lucidity, let $B \coloneqq \max_{v \in V} b(v)$.

Theorem 15.20. *The Col-BM on simple graphs $G = (V, E)$ with bounded treewidth $tw(G) < W$ is $\mathcal{XP}$ with respect to the number of colors q and the width bound W, and can be solved in $\mathcal{O}(|V| \cdot 2^{W^2 + W(q-1)} \cdot B^W \cdot \max\{2^{W+q}, B^W\})$ time.*

Proof. The correctness of the dynamic program follows from Lemma 15.19 and the label cost computations in Lemmas 15.15, 15.16, 15.17, and 15.18.

Concerning the algorithm's running time, recall that a nice tree decomposition $(T, \mathcal{X})$ of G with $\mathcal{O}(|V|)$ nodes can be computed in linear time. For each $t \in V(T)$, the number of labels $|\mathcal{L}^t|$ we have to consider at t is in

$$\mathcal{O}\left(2^{|E[X_t]|} \cdot 2^{q|X_t|} \cdot B^{|X_t|}\right) \subseteq \mathcal{O}\left(2^{W^2 - W} \cdot 2^{qW} \cdot B^W\right).$$

The computation of label costs for leaves and introduce nodes can be done in $\mathcal{O}(1)$ time.

For labels $x \in \mathcal{L}^t$ at forget node $t \in V(T)$ with child node $\ell \in V(T)$, we have to compare the label costs of $\left|\mathcal{L}^\ell(x)\right| = 2^{|U|}2^q$ labels. For simple graphs, $|U| \leq W$ and thus, the computation of label costs for forget nodes is in $\mathcal{O}(2^W 2^q)$ time.

For labels $x = (m, F, \beta) \in \mathcal{L}^t$ at a join node $t \in V(T)$ with child nodes $\ell, u \in V(T)$, $\left|\mathcal{L}^{\ell,u}(x)\right| = \Pi_{v \in X_t}(\beta_v + 1) \leq (B+1)^{|X_t|}$. Consequently, the computation of label costs for join nodes is in $\mathcal{O}(B^W)$ time.

In conclusion, the computation of label costs can be performed in $\mathcal{O}(\max\{2^{W+q}, B^W\})$ time. This results in a total running time in

$$\mathcal{O}(|V| \cdot 2^{W^2+W(q-1)} \cdot B^W \cdot \max\{2^{W+q}, B^W\}).$$

□

Corollary 15.21. *The Col-BM on simple graphs $G = (V, E)$ with treewidth $tw(G) < W$ is $\mathcal{FPT}$ with respect to the number of colors q, the width bound W, and the maximum b-value B.*

We note that for all Col-BM instances $B \leq |E|$ and thus, for fixed q and W, our dynamic program runs in $\mathcal{O}(|V| \cdot B^{2W})$ time, i.e., is polynomial. For trees, which are simple graphs with treewidth 1, the running time obtained from Theorem 15.20 coincides with the one from Corollary 15.14.

As soon as we drop the width bound W, we obtain the Col-BM on general graphs with a fixed number of colors which is strongly-$\mathcal{NP}$ hard by Theorem 15.2, even for $B = 2$. The complexity of the Col-BM on simple graphs $G = (V, E)$ with bounded treewidth $tw(G) < W$ and an arbitrary number of colors q remains open.

16 Discussion and Conclusion

We studied the multi budgeted matching problem and the minimum color-degree perfect b-matching problem – two specialized matching problems that originated in applications for the vehicle routing and staff assignment for MMUs. Using a reduction from 3-SAT, we showed the strong $\mathcal{NP}$-hardness of the mBM on paths and investigated the pseudo-polynomial solvability of the problem on special graph classes, namely series-parallel graphs, trees, and graphs of bounded treewidth. Regarding series-parallel graphs, we presented a dynamic program exploiting their representability in the form of decomposition trees. For trees, we used a simple graph transformation to reduce the problem to the former class of series-parallel graphs. Finally, for graphs of bounded treewidth, we suggested a dynamic program on nice tree decompositions that updates labels in a bottom-up fashion. All dynamic programs have pseudo-polynomial running time for a fixed number of budget constraints.

From an application-oriented point of view, the dynamic program for the mBM on graphs of bounded treewidth can be used to solve the vehicle routing problem for MMUs with multiple depots. However, as the graphs that derive from this application are mostly complete bipartite graphs $G = (V_A \cup V_B, E)$ which have treewidth $tw(G) = \min\{|V_A|, |V_B|\}$, the dynamic program is unlikely to outperform the binary programming formulation. Therefore, future work should focus on the mBM restricted to instances as they derive from the aforementioned MMU routing application. Moreover, it should be studied whether the multi-budgeted matching problem with fixed k on general graphs is strongly $\mathcal{NP}$-hard, or whether we can obtain a pseudo-polynomial algorithm. Even for the budgeted matching problem with a single budget constraint (BM) this question is still open. In 2011, Berger et al. (2011) conjectured that the BM is not strongly $\mathcal{NP}$-hard, i.e., that there exists a pseudo-polynomial algorithm for the BM. Until now, this conjecture could be neither proved nor disproved. The reduction presented in this thesis relies on a variant of the 3-SAT problem and requires one budget constraint per clause. However, for a fixed number of clauses 3-SAT is polynomial-time solvable by simple enumeration and thus every attempt at proving strong $\mathcal{NP}$-hardness has to be based on a new construction.

Considering the Col-BM, we proved the problem's strong $\mathcal{NP}$-hardness as well as its inapproximability for all constant approximation factors $1 < \alpha < 2$ on bipartite graphs with two colors. Subsequently, we identified a class of two-colored complete bipartite graphs on which we can solve the Col-BM in quadratic time and proposed polynomial-time dynamic programs solving the Col-BM with a fixed number of colors on series-parallel graphs and simple graphs with bounded treewidth.

Recalling our original application of the Col-BM for the staff assignment of MMUs, all our dynamic programs are likely to perform well as the number of different types of vehicles which corresponds to the number of colors in the edge coloring is likely to be relatively small. However, we must note that the computational performance also depends on the structure of the input graph which represents the possible assignments between physicians and MMU sessions. Future work on the Col-BM should include the generalization of the results for complete bipartite graphs to more colors as well as to more general b-values. Moreover, it would be interesting to investigate the complexity of the Col-BM on series-parallel graphs and graphs of bounded treewidth when the number of colors is not fixed. It could be further examined, how special structures in the edge coloring can be exploited. Finally, it might be possible to devise general exact algorithms and heuristics for the Col-BM by exploiting the structures of the underlying polytope.

As an outlook on how our results can be further applied in practice, we would like to point out that there is an ongoing research project that investigates the Col-BM for the patient-to-room assignment in hospitals. Presuming that patients of two genders have to be assigned to identical double rooms such that rooms are not gender-mixed, the patient-to-room assignment can be solved in polynomial time using the complete characterization from Section 15.2.

Bibliography

Aguwa, C., D. Egeonu, E.-E. Etu, J. Emakhu, O. Osoba, and L. Monplaisir (2018). "Multi-Criteria Allocation Configuration Problem for Mobile Health Clinics". In: *Proceedings of the 2018 IISE Annual Conference*, pp. 2080–2085 (cit. on p. 83).

Ahmadi-Javid, A., P. Seyedi, and S. S. Syam (2017). "A survey of healthcare facility location". In: *Computers & Operations Research* 79, pp. 223–263. DOI: 10.1016/j.cor.2016.05.018 (cit. on p. 82).

Ahmed, F., J. P. Dickerson, and M. Fuge (2017). "Diverse Weighted Bipartite b-Matching". In: *Proceedings of IJCAI-17*, pp. 35–41. DOI: 10.24963/ijcai.2017/6 (cit. on p. 153).

Aho (2006). "Bei Arztbesuchen sind die Deutschen Weltmeister". In: *Ärztliche Praxis*, p. 18 (cit. on p. 60).

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin (1993). *Network Flows. Theory, Algorithms, and Applications*. River: Prentice-Hall, Inc. (cit. on p. 93).

Alemayehu, B. and K. E. Warner (2004). "The Lifetime Distribution of Health Care Costs". In: *Health Services Research* 39.3, pp. 627–642. DOI: 10.1111/j.1475-6773.2004.00248.x (cit. on pp. 1, 77).

Alibrahim, A. and S. Wu (2018). "An agent-based simulation model of patient choice of health care providers in accountable care organizations". In: *Health Care Management Science* 21.1, pp. 131–143. DOI: 10.1007/s10729-014-9279-x (cit. on p. 18).

Amaldi, E., G. Galbiati, and F. Maffioli (2011). "On minimum reload cost paths, tours, and flows". In: *Networks* 57.3, pp. 254–260. DOI: 10.1002/net.20423 (cit. on p. 152).

American Academy of Family Physicians (2019). *Primary care*. URL: https://www.aafp.org/about/policies/all/primary-care.html (visited on Oct. 18, 2019) (cit. on p. 1).

Anapolska, M., C. Büsing, and M. Comis (2018). "Minimum color-degree perfect b-matchings". In: *16th Cologne-Twente Workshop on Graphs and Combinatorial Optimization*, pp. 13–16. eprint: http://ctw18.lipn.univ-paris13.fr/CTW18_Proceedings.pdf (cit. on pp. 153, 234).

Anapolska, M., C. Büsing, M. Comis, and T. Krabs (2021). "Minimum color-degree perfect b-matchings". In: *Networks* 77.4, pp. 477–494. DOI: 10.1002/net.21974 (cit. on pp. 153, 234).

Barahona, F. and W. R. Pulleyblank (1987). "Exact arborescences, matchings and cycles". In: *Discrete Applied Mathematics* 16.2, pp. 91–99. DOI: 10.1016/0166-218X(87)90067-9 (cit. on p. 151).

Barnes, S., B. Golden, and S. Price (2013). "Applications of Agent-Based Modeling and Simulation to Healthcare Operations Management". In: *Handbook of Healthcare Operations Management: Methods and Applications*. New York: Springer, pp. 45–74. DOI: 10.1007/978-1-4614-5885-2_3 (cit. on pp. 18, 21).

Baste, J., D. Gözüpek, M. Shalom, and D. M. Thilikos (2019). "Minimum Reload Cost Graph Factors". In: *SOFSEM 2019: Theory and Practice of Computer Science*. Ed. by B. Catania, R. Královič, J. Nawrocki, and G. Pighizzini. Cham: Springer International Publishing, pp. 67–80. DOI: doi.org/10.1007/978-3-030-10801-4_7 (cit. on p. 152).

Baumeister, R., E. Bratslavsky, C. Finkenauer, and K. Vohs (2001). “Bad is stronger than good”. In: *Review of General Psychology* 5.4, pp. 323–370. DOI: 10.1037/1089-2680.5.4.323 (cit. on p. 50).

Bechtold, S., B. Sommer, O. Pötzsch, and F. Burg (2019). *Bevölkerung im Wandel – Annahmen und Ergebnisse der 14. koordinierten Bevölkerungsvorausberechnung*. Statistisches Bundesamt, Wiesbaden. eprint: https://www.destatis.de/DE/Presse/Pressekonferenzen/2019/Bevoelkerung/pressebroschuere-bevoelkerung.pdf (cit. on p. 2).

Ben-Tal, A. and A. Nemirovski (2000). “Robust solutions of uncertain linear programming problems contaminated with uncertain data”. In: *Mathematical Programming* 88, pp. 411–421. DOI: 10.1007/PL00011380 (cit. on p. 95).

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009). *Robust optimization*. Vol. 28. Princeton: Princeton University Press (cit. on pp. 95, 97).

Beraldi, P. and A. Ruszczyński (2002). “The Probabilistic Set-Covering Problem”. In: *Operations Research* 50.6, pp. 956–967. DOI: 10.1287/opre.50.6.956.345 (cit. on p. 83).

Berger, A., V. Bonifaci, F. Grandoni, and G. Schäfer (2011). “Budgeted matching and budgeted matroid intersection via the gasoline puzzle”. In: *Mathematical Programming* 128.1, pp. 355–372. DOI: 10.1007/s10107-009-0307-4 (cit. on pp. 151, 199).

Berman, P., M. Karpinski, and A. D. Scott (2003). “Approximation Hardness of Short Symmetric Instances of MAX-3SAT”. In: *Electronic Colloquium on Computational Complexity (ECCC)*. eprint: https://eccc.weizmann.ac.il/report/2003/049/download (cit. on pp. 122, 178).

Bertsimas, D. and M. Sim (2003). “Robust discrete optimization and network flows”. In: *Mathematical Programming* 98.1, pp. 49–71. DOI: 10.1007/s10107-003-0396-4 (cit. on p. 97).

Bertsimas, D. and M. Sim (2004). “The Price of Robustness”. In: *Operartions Research* 52.1, pp. 35–53. DOI: 10.1287/opre.1030.0065 (cit. on pp. 97, 99, 130).

Bitton, A., H. L. Ratcliffe, J. H. Veillard, D. H. Kress, S. Barkley, M. Kimball, F. Secci, E. Wong, L. Basu, C. Taylor, et al. (2017). “Primary health care as a foundation for strengthening health systems in low-and middle-income countries”. In: *Journal of general internal medicine* 32.5, pp. 566–571. DOI: 10.1007/s11606-016-3898-5 (cit. on p. 1).

Bodenheimer, T. S. (1969). “Mobile Units: A Solution to the Rural Health Problem?” In: *Medical Care* 7.2, pp. 144–154. eprint: https://www.jstor.org/stable/3762568 (cit. on pp. 2, 77).

Bodlaender, H. L. (1996). “A Linear-Time Algorithm for Finding Tree-Decompositions of Small Treewidth”. In: *SIAM Journal on Computing* 25.6, pp. 1305–1317. DOI: 10.1137/S0097539793251219 (cit. on pp. 163, 175).

Bodlaender, H. L. (1988). “Dynamic programming on graphs with bounded treewidth”. In: *Automata, Languages and Programming, ICALP 1988*. Ed. by T. Lepistö and A. Salomaa. Vol. 317. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 105–118. DOI: 10.1007/3-540-19488-6_110 (cit. on p. 163).

Bodlaender, H. L. and A. M. C. A. Koster (May 2008). “Combinatorial Optimization on Graphs of Bounded Treewidth”. In: *The Computer Journal* 51.3, pp. 255–269. DOI: 10.1093/comjnl/bxm037 (cit. on p. 163).

Boeing, G. (2017). “OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks”. In: *Computers, Environment and Urban Systems* 65, pp. 126–139. DOI: 10.1016/j.compenvurbsys.2017.05.004 (cit. on p. 127).

Boscoe, F. P., K. A. Henry, and M. S. Zdeb (2012). “A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals”. In: *The Professional Geographer* 64.2, pp. 188–196. DOI: 10.1080/00330124.2011.583586 (cit. on p. 46).

Brailsford, S. C., P. R. Harper, B. Patel, and M. Pitt (Sept. 2009). “An analysis of the academic literature on simulation and modelling in health care”. In: *Journal of Simulation* 3.3, pp. 130–140. DOI: 10.1057/jos.2009.10 (cit. on p. 20).

Brailsford, S. C. (2008). “System dynamics: What’s in it for healthcare simulation modelers”. In: *Proceedings of the 2008 Winter Simulation Conference*, pp. 1478–1483. DOI: 10.1109/WSC.2008.4736227 (cit. on p. 20).

Brailsford, S. C., S. M. Desai, and J. Viana (2010). “Towards the holy grail: Combining system dynamics and discrete-event simulation in healthcare”. In: *Proceedings of the 2010 Winter Simulation Conference*, pp. 2293–2303. DOI: 10.1109/WSC.2010.5678927 (cit. on p. 21).

Brailsford, S. C., T. Eldabi, M. Kunc, N. Mustafee, and A. F. Osorio (2019). “Hybrid simulation modelling in operational research: A state-of-the-art review”. In: *European Journal of Operational Research* 278.3, pp. 721–737. DOI: 10.1016/j.ejor.2018.10.025 (cit. on p. 21).

Bundestag (2019). “Gesetz für schnellere Termine und bessere Versorgung (Terminservice- und Versorgungsgesetz– TSVG)”. In: *Bundesgesetzblatt* 1.18, pp. 646–691. eprint: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/T/TSVG_BGBL.pdf (cit. on p. 72).

Bureau of Health Workforce, Health Resources and Services Administration (HRSA) (2020). *Designated Health Professional Shortage Areas Statistics*. U.S. Department of Health & Human Services, pp. 1–15. eprint: https://data.hrsa.gov/Default/GenerateHPSAQuarterlyReport (cit. on p. 18).

Büsing, C. and M. Comis (2018a). “Budgeted Colored Matching Problems”. In: *Electronic Notes in Discrete Mathematics* 64. 8th International Network Optimization Conference - INOC 2017, pp. 245–254. DOI: 10.1016/j.endm.2018.01.026 (cit. on pp. 84, 152, 233).

Büsing, C. and M. Comis (2018b). “Multi-budgeted matching problems”. In: *Networks* 72.1, pp. 25–41. DOI: 10.1002/net.21802 (cit. on pp. 153, 234).

Büsing, C., M. Comis, E. Schmidt, and M. Streicher (2021). “Robust strategic planning for mobile medical units with steerable and unsteerable demands”. In: *European Journal of Operational Research* 295.1, pp. 34–50. DOI: 10.1016/j.ejor.2021.02.037 (cit. on pp. 84, 233).

Büsing, C., M. Comis, and S. Schmitz (2020a). “Robust Mask-Based Planning for Appointment Scheduling in Primary Care Practices”. working paper, unpublished (cit. on p. 73).

Büsing, C., A. M. C. A. Koster, S. Kirchner, and A. Thome (2016). “The budgeted minimum cost flow problem with unit upgrading cost”. In: *Networks* 69.1, pp. 67–82. DOI: 10.1002/net.21724 (cit. on p. 163).

Büsing, C., S. Schmitz, M. Anapolska, S. Theis, M. Wille, C. Brandl, V. Nitsch, and A. Mertens (2020b). “Agent-Based Simulation of Medical Care Processes in Rural Areas with the Aid of Current Data on ICT Usage Readiness Among Elderly Patients”. In: *Human Aspects of IT for the Aged Population. Healthy and Active Aging - 6th International Conference, ITAP 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part II*. Ed. by Q. Gao and J. Zhou. Vol. 12208. Lecture Notes in Computer Science. Springer, pp. 3–12. DOI: 10.1007/978-3-030-50249-2_1 (cit. on p. 72).

Calik, H., M. Labbé, and H. Yaman (2015). “p-Center Problems”. In: *Location Science*. Ed. by G. Laporte, S. Nickel, and F. Saldanha da Gama. Cham: Springer International Publishing, pp. 79–92. DOI: 10.1007/978-3-319-13111-5_4 (cit. on p. 83).

Camerini, P., G. Galbiati, and F. Maffioli (1992). "Random pseudo-polynomial algorithms for exact matroid problems". In: *J. Algorithms* 13.2, pp. 258–273. DOI: 10.1016/0196-6774(92)90018-8 (cit. on p. 151).

Cánovas, L., S. García Quiles, M. Labbé, and A. Marín (2007). "A strengthened formulation for the simple plant location problem with order". In: *Operations Research Letters* 35, pp. 141–150. DOI: 10.1016/j.orl.2006.01.012 (cit. on p. 83).

Caprara, A., P. Toth, and M. Fischetti (2000). "Algorithms for the set covering problem". In: *Annals of Operations Research* 98.1-4, pp. 353–371. DOI: 10.1023/A:1019225027893 (cit. on p. 82).

Carrabs, F., R. Cerulli, and M. Gentili (2009). "The labeled maximum matching problem". In: *Computers & Operations Research* 36.6, pp. 1859–1871. DOI: 10.1016/j.cor.2008.05.012 (cit. on p. 152).

Cayirli, T. and E. Veral (2003). "Outpatient scheduling in health care: A review of literature". In: *Production and Operations Management* 12.4, pp. 519–549. DOI: 10.1111/j.1937-5956.2003.tb00218.x (cit. on pp. 41, 52, 56).

Cayirli, T., E. Veral, and H. Rosen (Feb. 2006). "Designing appointment scheduling systems for ambulatory care services". In: *Health Care Management Science* 9.1, pp. 47–58. DOI: 10.1007/s10729-006-6279-5 (cit. on pp. 21, 22, 41).

Chekuri, C., J. Vondrák, and R. Zenklusen (2011). "Multi-budgeted Matchings and Matroid Intersection via Dependent Rounding". In: *Proceedings of the 22 Annual ACM-SIAM Symposium on Discrete Algorithms*. San Francisco, California: Society for Industrial and Applied Mathematics, pp. 1080–1097. DOI: 10.1137/1.9781611973082.82 (cit. on p. 152).

Comis, M., C. Cleophas, and C. Büsing (2021). "Patients, primary care, and policy: Agent-based simulation modeling for health care decision support". In: *Health Care Management Science*. DOI: 10.1007/s10729-021-09556-2 (cit. on pp. 22, 24, 233).

Cook, S. A. (1971). "The Complexity of Theorem-Proving Procedures". In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing*. STOC '71. Shaker Heights, Ohio, USA: Association for Computing Machinery, pp. 151–158. DOI: 10.1145/800157.805047 (cit. on p. 11).

Cox, T. F., J. P. Birchall, and H. Wong (1985). "Optimising the queuing system for an ear, nose and throat outpatient clinic". In: *Journal of Applied Statistics* 12.2, pp. 113–126. DOI: 10.1080/02664768500000017 (cit. on p. 52).

Daley, D. and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. 2nd ed. New York: Springer, pp. 1–471. DOI: 10.1007/b97277 (cit. on p. 40).

Darmann, A., U. Pferschy, J. Schauer, and G. J. Woeginger (2011). "Paths, trees and matchings under disjunctive constraints". In: *Discrete Applied Mathematics* 159.16, pp. 1726–1735. DOI: 10.1016/j.dam.2010.12.016 (cit. on p. 122).

Devlin, K. J. (2002). *The millennium problems: the seven greatest unsolved mathematical puzzles of our time*. New York: Basic books (cit. on p. 11).

Dhamdhere, K., V. Goyal, R. Ravi, and M. Singh (2005). "How to pay, come what may: approximation algorithms for demand-robust covering problems". In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pp. 367–376. DOI: 10.1109/SFCS.2005.42 (cit. on p. 82).

Dodd, R., ed. (2005). *Health and the millennium development goals*. World Health Organization (cit. on p. 1).

Doerner, K., A. Focke, and W. J. Gutjahr (2007). “Multicriteria tour planning for mobile healthcare facilities in a developing country”. In: *European Journal of Operational Research* 179.3, pp. 1078–1096. DOI: 10.1016/j.ejor.2005.10.067 (cit. on pp. 81, 85).

Downey, R. G. and M. R. Fellows (1999). *Parameterized Complexity*. New York: Springer (cit. on p. 12).

Dye, C., J. C. Reeder, and R. F. Terry (2013). “Research for Universal Health Coverage”. In: *Science Translational Medicine* 5.199, 199ed13. DOI: 10.1126/scitranslmed.3006971 (cit. on p. 1).

Farahani, R. Z., N. Asgari, N. Heidari, M. Hosseininia, and M. Goh (2012). “Covering problems in facility location: A review”. In: *Computers & Industrial Engineering* 62.1, pp. 368–407. DOI: 10.1016/j.cie.2011.08.020 (cit. on p. 82).

Feige, U., K. Jain, M. Mahdian, and V. Mirrokni (2007). “Robust Combinatorial Optimization with Exponential Scenarios”. In: *Integer Programming and Combinatorial Optimization*. Ed. by D. P. Fischetti Matteo and Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 439–453. DOI: 10.1007/978-3-540-72792-7_33 (cit. on p. 82).

Fetter, R. B. and J. M. D. Thompson (1966). “Patients’ waiting time and doctors’ idle time in the outpatient setting”. In: *Health services research* 1.1, pp. 66–90 (cit. on p. 41).

Fischetti, M. and M. Monaci (2012). “Cutting plane versus compact formulations for uncertain (integer) linear programs”. In: *Mathematical Programming Computation* 4.3, pp. 239–273. DOI: 10.1007/s12532-012-0039-y (cit. on p. 83).

Fone, D., S. Hollinghurst, J. Temple, A. Round, N. Lester, A. Weightman, K. Roberts, E. Coyle, G. Bevan, and S. Palmer (2003). “Systematic review of the use and value of computer simulation modelling in population health and health care delivery”. In: *Journal of Public Health Medicine* 25.4, pp. 325–335. DOI: 10.1093/pubmed/fdg075 (cit. on pp. 2, 20, 85).

Fujita, S. and C. Magnant (2011). “Properly colored paths and cycles”. In: *Discrete Applied Mathematics* 159.14, pp. 1391–1397. DOI: 10.1016/j.dam.2011.06.005 (cit. on p. 178).

Gabrel, V., C. Murat, and A. Thiele (2014). “Recent advances in robust optimization: An overview”. In: *European Journal of Operational Research* 235.3, pp. 471–483. DOI: 10.1016/j.ejor.2013.09.036 (cit. on p. 95).

Gale, D. and L. S. Shapley (1962). “College Admissions and the Stability of Marriage”. In: *The American Mathematical Monthly* 69.1, pp. 9–15. DOI: 10.1080/00029890.1962.11989827 (cit. on p. 149).

García, S. and A. Marín (2015). “Covering location problems”. In: *Location science*. Springer, pp. 93–114. DOI: 10.1007/978-3-319-13111-5_5 (cit. on p. 82).

Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman & Co. (cit. on pp. 9, 101, 107, 152).

Gemeinsamer Bundesausschuss (2012). “Bedarfsplanungs-Richtlinie”. In: *Bundesanzeiger, Bekanntmachung BAnz AT 31.12.2012 B7*. eprint: https://www.g-ba.de/downloads/62-492-2022/BPL-RL_2019-12-05_iK-2019-12-21.pdf (cit. on pp. 17, 18).

Geoffroy, E., A. Harries, K. Bissell, E. Schell, A. Bvumbwe, K. Tayler-Smith, and W. Kizito (2014). “Bringing care to the community: expanding access to health care in rural Malawi through mobile health clinics”. In: *Public Health Action* 4.4, pp. 252–258. DOI: 10.5588/pha.14.0064 (cit. on p. 78).

Giachetti, R. E., E. A. Centeno, M. A. Centeno, and R. Sundaram (2005). “Assessing the viability of an open access policy in an outpatient clinic: a discrete-event and continuous simulation modeling approach”. In: *Proceedings of the 2005 Winter Simulation Conference*, pp. 2246–2255. DOI: 10.1109/WSC.2005.1574513 (cit. on pp. 21, 22).

Gilbert, N. (2008). *Agent-based models*. Quantitative Applications in the Social Sciences 153. Thousand Oaks: Sage Publications (cit. on pp. 18, 20).

Gourvès, L., A. Lyra, C. Martinhon, and J. Monnot (2009). "The Minimum Reload s-t Path/Trail/Walk Problems". In: *SOFSEM 2009: Theory and Practice of Computer Science*. Ed. by M. Nielsen, A. Kučera, P. B. Miltersen, C. Palamidessi, P. Tůma, and F. Valencia. Berlin, Heidelberg: Springer, pp. 621–632. DOI: 10.1007/978-3-540-95891-8_55 (cit. on p. 152).

Grandoni, F., R. Ravi, M. Singh, and R. Zenklusen (2014). "New approaches to multi-objective optimization". In: *Mathematical Programming* 146.1, pp. 525–554. DOI: 10.1007/s10107-013-0703-7 (cit. on p. 157).

Grandoni, F. and R. Zenklusen (2010). "Approximation Schemes for Multi-Budgeted Independence Systems". In: *Algorithms – ESA 2010: 18th Annual European Symposium. Proceedings, Part I*. Berlin, Heidelberg: Springer, pp. 536–548. DOI: 10.1007/978-3-642-15775-2_46 (cit. on p. 151).

Grimm, V., U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback (2010). "The ODD protocol: A review and first update". In: *Ecological Modelling* 221.23, pp. 2760–2768. DOI: 10.1016/j.ecolmodel.2010.08.019 (cit. on pp. 23, 25).

Grobe, T., H. Dörning, and F. Schwartz (2011). *BARMER GEK Arztreport 2011*. St. Augustin: Asgard-Verlag. eprint: https://www.barmer.de/blob/36506/d5630a0f349e388b65fd28ad616b7257/data/arztreport-2011-pdf.pdf (cit. on pp. 56, 59).

Gupta, A., V. Nagarajan, and R. Ravi (2014). "Thresholded covering algorithms for robust and max–min optimization". In: *Mathematical Programming* 146.1-2, pp. 583–615. DOI: 10.1007/s10107-013-0705-5 (cit. on p. 82).

Gupta, D. and B. Denton (2008). "Appointment scheduling in health care: Challenges and opportunities". In: *IIE Transactions* 40.9, pp. 800–819. DOI: 10.1080/07408170802165880 (cit. on p. 34).

Hachicha, M., M. J. Hodgson, G. Laporte, and F. Semet (2000). "Heuristics for the multi-vehicle covering tour problem". In: *Computers & Operations Research* 27.1, pp. 29–42. DOI: 10.1016/S0305-0548(99)00006-4 (cit. on p. 81).

Hakimi, S. L. (1965). "Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems". In: *Operations Research* 13.3, pp. 462–475. DOI: 10.1287/opre.13.3.462 (cit. on p. 83).

Hamrock, E., K. Paige, J. Parks, J. Scheulen, and S. Levin (2013). "Discrete event simulation for healthcare organizations: a tool for decision making". In: *Journal of Healthcare Management* 58.2, pp. 110–124. DOI: 10.1097/00115514-201303000-00007 (cit. on p. 20).

Hanjoul, P. and D. Peeters (1987). "A facility location problem with clients' preference orderings". In: *Regional Science and Urban Economics* 17.3, pp. 451–473. DOI: 10.1016/0166-0462(87)90011-1 (cit. on p. 83).

Heller, I. and C. Tompkins (1956). "An extension of a theorem of Dantzig's". In: *Linear inequalities and related systems* 38, pp. 247–254. DOI: 10.1515/9781400881987-015 (cit. on p. 91).

Hill, C., B. Powers, S. Jain, J. Bennet, A. Vavasis, and N. Oriol (2014). "Mobile Health Clinics in the Era of Reform". In: *The American journal of managed care* 20, pp. 261–264. eprint: http://ajmc.s3.amazonaws.com/_media/_pdf/AJMC_03_14_Hill_261to64.pdf (cit. on pp. 77, 78).

Hochbaum, D. S. and D. B. Shmoys (1985). "A Best Possible Heuristic for the k-Center Problem". In: *Mathematics of Operations Research* 10.2, pp. 180–184. DOI: 10.1287/moor.10.2.180 (cit. on p. 83).

Hodgson, M. J., G. Laporte, and F. Semet (1998). “A Covering Tour Model for Planning Mobile Health Care Facilities in Suhum District, Ghana”. In: *Journal of Regional Science* 38.4, pp. 621–638. DOI: 10.1111/0022-4146.00113 (cit. on p. 81).

Homa, L., J. Rose, P. S. Hovmand, et al. (2015). “A participatory model of the paradox of primary care”. In: *Annals of Family Medicine* 13.5, pp. 456–465. DOI: 10.1370/afm.1841 (cit. on pp. 22, 23).

Homer, J. B. and G. B. Hirsch (2006). “System dynamics modeling for public health: background and opportunities”. In: *American Journal of Public Health* 96.3, pp. 452–458. DOI: 10.2105/AJPH.2005.062059 (cit. on p. 20).

Hsu, W.-L. and G. L. Nemhauser (1979). “Easy and hard bottleneck location problems”. In: *Discrete Applied Mathematics* 1.3, pp. 209–215. DOI: https://doi.org/10.1016/0166-218X(79)90044-1 (cit. on p. 83).

Huigen, P. P., A. H. H. M. Kempers-Warderdam, and C. Volkers (1986). “Demographic changes and service provision in rural areas in the Netherlands”. In: *Espace Populations Sociétés* 4.3, pp. 55–62. eprint: https://www.persee.fr/doc/espos_0755-7809_1986_num_4_3_1160 (cit. on p. 2).

IBM (2018). *IBM CPLEX Optimization Studio 12.8*. URL: http://www.cplex.com/ (visited on Apr. 18, 2020) (cit. on pp. 129, 135, 137).

Information und Technik Nordrhein-Westfalen (2019). *Bevölkerungsentwicklung 2018 – 2060 nach Altersgruppen am 1. Januar*. https://www.it.nrw/node/971/pdf. Accessed: 2019-10-18 (cit. on pp. 55, 64).

Information und Technik Nordrhein-Westfalen (2016). *Zensus 2011: Vielfältiges Deutschland*. Statistische Ämter des Bundes und der Länder. eprint: https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2016_12_NRW_Zensus_Vielfalt.pdf (cit. on pp. 55–58, 127).

International Air Transport Association (2019). *Worldwide Slot Guidelines*. 10th ed. eprint: https://www.iata.org/contentassets/4ede2aabfcc14a55919e468054d714fe/wsg-edition-10-english-version.pdf (cit. on p. 177).

Irving, G., A. L. Neves, H. Dambha-Miller, A. Oishi, H. Tagashira, A. Verho, and J. Holden (2017). “International variations in primary care physician consultation time: a systematic review of 67 countries”. In: *BMJ Open* 7.10. DOI: 10.1136/bmjopen-2017-017902 (cit. on pp. 125, 127).

Jacob, R., J. Kopp, and S. Schultz (2015). *Berufsmonitoring Medizinstudenten 2014. Ergebnisse einer bundesweiten Befragung*. Kassenärztliche Bundesvereinigung, pp. 1–100. eprint: https://www.kbv.de/media/sp/2015-04-08_Berufsmonitoring_2014_web.pdf (cit. on pp. 2, 55).

Jacobson, S. H., S. N. Hall, and J. R. Swisher (2006). “Discrete-Event Simulation of Health Care Systems”. In: *Patient Flow: Reducing Delay in Healthcare Delivery*. Ed. by R. W. Hall. Vol. 91. International Series in Operations Research & Management Science. Boston: Springer, pp. 211–252. DOI: 10.1007/978-0-387-33636-7_8 (cit. on p. 21).

Kano, M. and X. Li (Sept. 2008). “Monochromatic and Heterochromatic Subgraphs in Edge-Colored Graphs - A Survey”. In: *Graphs and Combinatorics* 24.4, pp. 237–263. DOI: 10.1007/s00373-008-0789-5 (cit. on p. 181).

Kasperski, A. (2008). *Discrete Optimization with Interval Data: Minmax Regret and Fuzzy Approach*. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer (cit. on p. 97).

Kasperski, A. and P. Zieliński (2016). "Robust Discrete Optimization Under Discrete and Interval Uncertainty: A Survey". In: *Robustness Analysis in Decision Aiding, Optimization, and Analytics*. Ed. by M. Doumpos, C. Zopounidis, and E. Grigoroudis. Vol. 241. International Series in Operations Research & Management Science. Springer, pp. 113–143. DOI: 10.1007/978-3-319-33121-8_6 (cit. on p. 97).

Kassenärtzliche Vereinigung Nordrhein (2018). " Die 100 häufigsten ICD-10-Schlüssel und Kurztexte (nach Fachgruppen) - 4. Quartal 2018." In: pp. 1–125. eprint: https://www.kvno.de/fileadmin/shared/pdf/online/verordnungen/morbiditaetsstatistik/100icd_18-4.pdf (cit. on p. 59).

Kassenärtzliche Vereinigung Nordrhein (2019). *Suche nach Ärzten und Psychotherapeuten in Nordrhein*. URL: https://www.kvno.de/20patienten/10arztsuche/ (visited on Oct. 18, 2019) (cit. on pp. 55, 56, 125).

Kassenärztliche Bundesvereinigung (2017). *Statistische Informationen aus dem Bundesarztregister*. Accessed: 2020-06-15. eprint: https://www.kbv.de/media/sp/2017_12_31_BAR_Statistik.pdf (cit. on p. 2).

Khanna, A. B. and S. A. Narula (2017). "Mobile Medical Units–Can They Improve the Quality of Health Services in Developing Countries?" In: *Journal of Health Management* 19.3, pp. 508–521. DOI: 10.1177/0972063417717900 (cit. on pp. 77, 78).

Khuller, S. and Y. J. Sussmann (2000). "The Capacitated K-Center Problem". In: *SIAM Journal on Discrete Mathematics* 13.3, pp. 403–418. DOI: 10.1137/S0895480197329776 (cit. on p. 83).

Kikuno, T., N. Yoshida, and Y. Kakuda (1983). "A linear algorithm for the domination number of a series-parallel graph". In: *Discrete Applied Mathematics* 5.3, pp. 299–311. DOI: 10.1016/0166-218X(83)90003-3 (cit. on p. 157).

Kim, S.-H., C. W. Chan, M. Olivares, and G. Escobar (2015). "ICU Admission Control: An Empirical Study of Capacity Allocation and Its Implication for Patient Outcomes". In: *Management Science* 61.1, pp. 19–38. DOI: 10.1287/mnsc.2014.2057 (cit. on pp. 52, 56).

Klassen, K. J. and T. R. Rohleder (1996). "Scheduling outpatient appointments in a dynamic environment". In: *Journal of Operations Management* 14.2, pp. 83–101. DOI: 10.1016/0272-6963(95)00044-5 (cit. on pp. 27, 52, 56, 77).

Kleijnen, J. P. (1995). "Verification and validation of simulation models". In: *European Journal of Operational Research* 82.1, pp. 145–162. DOI: 10.1016/0377-2217(94)00016-6 (cit. on p. 52).

Kloks, T. (1994). *Treewidth, Computations and Approximations*. Vol. 842. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer (cit. on pp. 163, 164, 175).

Korenblit, M. and V. E. Levit (2011). "On the Structure of Maximum Series Parallel Graphs". In: *AIP Conference Proceedings*. Vol. 1389, pp. 1932–1935. DOI: 10.1063/1.3636990 (cit. on p. 176).

Korte, B. and J. Vygen (2012). *Combinatorial Optimization: Theory and Algorithms*. Algorithms and Combinatorics. Berlin Heidelberg: Springer (cit. on pp. 7, 9, 13, 14).

Korzilius, H. (2019). "Terminservice- und Versorgungsgesetz: Bundestag billigt umstrittenes Gesetz". In: *Deutsche Ärzteblatt* 116.12, A555–A557 (cit. on p. 72).

Koster, A. M. C. A. (2017). *Algorithmic Graph Theory: How hard is your combinatorial optimization problem?* Lecture notes. University of Clemson (cit. on p. 12).

Kouvelis, P. and G. Yu (1996). *Robust Discrete Optimization and Its Applications*. Nonconvex Optimization and Its Applications. Boston: Springer (cit. on p. 97).

Kringos, D. S., W. G. Boerma, A. Hutchinson, and R. B. Saltman (May 2015). *Building primary care in a changing Europe: case studies*. World Health Organization, pp. 1–304. eprint: https://www.euro.who.int/__data/assets/pdf_file/0011/277940/Building-primary-care-changing-Europe-case-studies.pdf (cit. on p. 18).

Krumke, S. O., E. Schmidt, and M. Streicher (2019). "Robust multicovers with budgeted uncertainty". In: *European Journal of Operational Research* 274.3, pp. 845–857. DOI: 10.1016/j.ejor.2018.11.049 (cit. on pp. 82, 87, 90, 101).

Kuhn, H. W. (1955). "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2, pp. 83–97. DOI: 10.1002/nav.3800020109 (cit. on pp. 118, 149).

Le, V. B. and F. Pfender (2014). "Complexity results for rainbow matchings". In: *Theoretical Computer Science* 524, pp. 27–33. DOI: 10.1016/j.tcs.2013.12.013 (cit. on p. 152).

Lim, A., B. Rodrigues, F. Wang, and Z. Xu (2005). "k-Center problems with minimum coverage". In: *Theoretical Computer Science* 332.1, pp. 1–17. DOI: https://doi.org/10.1016/j.tcs.2004.08.010 (cit. on p. 83).

Liu, P. and S. Wu (2016). "An agent-based simulation model to study accountable care organizations". In: *Health Care Management Science* 19.1, pp. 89–101. DOI: 10.1007/s10729-014-9279-x (cit. on p. 18).

Lopes, M. A., Á. S. Almeida, and B. Almada-Lobo (Mar. 2018). "Forecasting the medical workforce: a stochastic agent-based simulation approach". In: *Health Care Management Science* 21.1, pp. 52–75. DOI: 10.1007/s10729-016-9379-x (cit. on p. 19).

Lutter, P., D. Degel, C. Büsing, A. M. C. A. Koster, and B. Werners (2017). "Improved handling of uncertainty and robustness in set covering problems". In: *European Journal of Operational Research* 263.1, pp. 35–49. DOI: 10.1016/j.ejor.2017.04.044 (cit. on p. 83).

Mann, E., B. Schuetz, and E. Rubin-Johnston (2010). *Remaking Primary Care. A Framework for the Future*. Cambridge: New England Healthcare Institute, pp. 1–47. eprint: https://www.nehi.net/writable/publication_files/file/remaking_primary_care_a_framework_for_the_future_final.pdf (cit. on pp. 1–3, 17, 77).

Maroulis, S., R. Guimerà, H. Petry, M. J. Stringer, L. M. Gomez, L. A. N. Amaral, and U. Wilensky (2010). "Complex Systems View of Educational Policy Research". In: *Science* 330.6000, pp. 38–39. DOI: 10.1126/science.1195153 (cit. on p. 18).

Mastrolilli, M. and G. Stamoulis (2014). "Bi-criteria and approximation algorithms for restricted matchings". In: *Theoretical Computer Science* 540-541. Combinatorial Optimization: Theory of algorithms and complexity, pp. 115–132. DOI: 10.1016/j.tcs.2013.11.027 (cit. on pp. 152, 153).

Mastrolilli, M. and G. Stamoulis (2012). "Constrained Matching Problems in Bipartite Graphs". In: *Combinatorial Optimization*. Ed. by A. R. Mahjoub, V. Markakis, I. Milis, and V. T. Paschos. Berlin, Heidelberg: Springer, pp. 344–355. DOI: 10.1007/978-3-642-32147-4_31 (cit. on p. 152).

Matchar, D. B., J. P. Ansah, P. Hovmand, and S. Bayer (Dec. 2016). "Simulation modeling for primary care planning in Singapore". In: *Proceedings of the 2016 Winter Simulation Conference*, pp. 2123–2134. DOI: 10.1109/WSC.2016.7822255 (cit. on pp. 22, 23).

Math Commons (2016). *The apache commons mathematics library*. URL: https://commons.apache.org/proper/commons-math/ (visited on Nov. 12, 2020) (cit. on p. 53).

Meng, Y., R. Davies, K. Hardy, and P. Hawkey (Mar. 2010). "An application of agent-based simulation to the management of hospital-acquired infection". In: *Journal of Simulation* 4.1, pp. 60–67. DOI: 10.1057/jos.2009.17 (cit. on p. 19).

Mihelič, J. and B. Robič (2003). “Approximation Algorithms for the k-center Problem: An Experimental Evaluation”. In: *Operations Research Proceedings 2002*. Ed. by U. Leopold-Wildburger, F. Rendl, and G. Wäscher. Berlin, Heidelberg: Springer, pp. 371–376. DOI: 10.1007/978-3-642-55537-4_60 (cit. on p. 83).

Mihelič, J. and B. Robič (2005). “Solving the k-center Problem Efficiently with a Dominating Set Algorithm”. In: *Journal of Computing and Information Technology* 13.3, pp. 225–234. DOI: 10.2498/cit.2005.03.05 (cit. on p. 145).

Monnot, J. (2005). “The labeled perfect matching in bipartite graphs”. In: *Information Processing Letters* 96.3, pp. 81–88. DOI: 10.1016/j.ipl.2005.06.009 (cit. on p. 152).

Monnot, J. and S. Toulouse (2007). “The path partition problem and related problems in bipartite graphs”. In: *Operations Research Letters* 35.5, pp. 677–684. DOI: 10.1016/j.orl.2006.12.004 (cit. on p. 180).

Murray, A. T., R. Davis, R. J. Stimson, and L. Ferreira (1998). “Public Transportation Access”. In: *Transportation Research Part D: Transport and Environment* 3.5, pp. 319–328. DOI: 10.1016/S1361-9209(98)00010-8 (cit. on p. 2).

Naji-Azimi, Z., J. Renaud, A. Ruiz, and M. Salari (2012). “A covering tour approach to the location of satellite distribution centers to supply humanitarian aid”. In: *European Journal of Operational Research* 222.3, pp. 596–605. DOI: 10.1016/j.ejor.2012.05.001 (cit. on p. 82).

Nemhauser, G. and L. Wolsey (2014a). “Integral Polyhedra”. In: *Integer and Combinatorial Optimization*. John Wiley & Sons, Ltd. Chap. III.1, pp. 533–607. DOI: 10.1002/9781118627372.ch14 (cit. on p. 99).

Nemhauser, G. and L. Wolsey (2014b). “Linear Programming”. In: *Integer and Combinatorial Optimization*. John Wiley & Sons, Ltd. Chap. I.2, pp. 27–49. DOI: 10.1002/9781118627372.ch2 (cit. on p. 100).

Nomikos, C., A. Pagourtzis, and S. Zachos (2007). “Randomized and Approximation Algorithms for Blue-Red Matching”. In: *Mathematical Foundations of Computer Science 2007*. Ed. by L. Kučera and A. Kučera. Berlin, Heidelberg: Springer, pp. 715–725. DOI: 10.1007/978-3-540-74456-6_63 (cit. on p. 152).

Northridge, M. E. and S. S. Metcalf (2016). “Enhancing implementation science by applying best principles of systems science”. In: *Health research policy and systems* 14.1, p. 74. DOI: 10.1186/s12961-016-0146-8 (cit. on p. 25).

Oh, H. J. A., A. Muriel, and H. Balasubramanian (Dec. 2014). “A user-friendly Excel simulation for scheduling in primary care practices”. In: *Proceedings of the 2014 Winter Simulation Conference*, pp. 1177–1185. DOI: 10.1109/WSC.2014.7019975 (cit. on p. 21).

OpenStreetMap contributors (2019). planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org (cit. on p. 127).

Oracle (2018). *Open Java Development Kit*. URL: https://openjdk.java.net/ (visited on Mar. 18, 2020) (cit. on pp. 10, 20, 52, 129, 135, 137).

Ozbaygin, G., H. Yaman, and O. E. Karasan (2016). “Time constrained maximal covering salesman problem with weighted demands and partial coverage”. In: *Computers & Operations Research* 76, pp. 226–237. DOI: 10.1016/j.cor.2016.06.019 (cit. on p. 81).

Patlolla, P., V. Gunupudi, A. R. Mikler, and R. T. Jacob (2006). “Agent-Based Simulation Tools in Computational Epidemiology”. In: *Innovative Internet Community Systems*. Ed. by T. Böhme, V. M. Larios Rosillo, H. Unger, and H. Unger. Berlin, Heidelberg: Springer, pp. 212–223. DOI: 10.1007/11553762_21 (cit. on p. 19).

Patro, B., R. Kumar, A. Goswami, B. Nongkynrih, and C. Pandav (2008). “Community Perception and Client Satisfaction about the Primary Health Care Services in an Urban Resettlement Colony of New Delhi”. In: *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine* 33, pp. 250–254. DOI: 10.4103/0970-0218.43232 (cit. on p. 78).

Pentico, D. W. (2007). “Assignment problems: A golden anniversary survey”. In: *European Journal of Operational Research* 176.2, pp. 774–793. DOI: https://doi.org/10.1016/j.ejor.2005.09.014 (cit. on pp. 149, 153).

Pereira, J. and I. Averbakh (2013). “The Robust Set Covering Problem with interval data”. In: *Annals of Operartions Research* 207, pp. 1–19. DOI: 10.1007/s10479-011-0876-5 (cit. on p. 82).

Pfaff, H., E. Neugebauer, G. Glaeske, M. Schrappe, M. Rothmund, and W. Schwartz (2017). *Lehrbuch Versorgungsforschung: Systematik - Methodik - Anwendung*. Stuttgart: Schattauer (cit. on pp. 2, 17).

Qu, X., Y. Peng, J. Shi, and L. LaGanga (2015). “An MDP model for walk-in patient admission management in primary care clinics”. In: *International Journal of Production Economics* 168, pp. 303–320. DOI: 10.1016/j.ijpe.2015.06.022 (cit. on p. 52).

Rigotti, N. and R. Wallace (2015). “Using agent-based models to address “wicked problems” like tobacco use: A report from the institute of medicine”. In: *Annals of Internal Medicine* 163.6, pp. 469–471. DOI: 10.7326/M15-1567 (cit. on p. 18).

Rising, E. J., R. Baron, and B. Averill (1973). “A Systems Analysis of a University-Health-Service Outpatient Clinic”. In: *Operations Research* 21.5, pp. 1030–1047. DOI: 10.1287/opre.21.5.1030 (cit. on p. 52).

Rittel, H. W. J. and M. M. Webber (June 1973). “Dilemmas in a general theory of planning”. In: *Policy Sciences* 4.2, pp. 155–169. DOI: 10.1007/BF01405730 (cit. on p. 18).

Robert Koch-Institut (2014). “Chronisches Kranksein”. In: *Faktenblatt zu GEDA 2012: Ergebnisse der Studie Gesundheit in Deutschland aktuell 2012*, pp. 1–4. eprint: https://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsF/Geda2012/chronisches_kranksein.pdf (cit. on pp. 56, 58).

Robertson, N. and P. Seymour (1986). “Graph minors. II. Algorithmic aspects of tree-width”. In: *Journal of Algorithms* 7.3, pp. 309–322. DOI: 10.1016/0196-6774(86)90023-4 (cit. on p. 163).

Rosenbrock, R. and T. Gerlinger (2014). *Gesundheitspolitik. Eine systematische Einführung*. Bern: Verlag Hans Huber (cit. on pp. 3, 17).

Rusu, I. (2008). “Maximum weight edge-constrained matchings”. In: *Discrete Applied Mathematics* 156.5, pp. 662–672. DOI: 10.1016/j.dam.2007.08.021 (cit. on p. 152).

Sargent, R. G. (Feb. 2013). “Verification and validation of simulation models”. In: *Journal of Simulation* 7.1, pp. 12–24. DOI: 10.1057/jos.2012.20 (cit. on p. 52).

Saxena, A., V. Goyal, and M. Lejeune (2010). “MIP Reformulations of the Probabilistic Set Covering Problem”. In: *Mathematical Programming* 121, pp. 1–21. DOI: 10.1007/s10107-008-0224-y (cit. on p. 83).

Schacht, M. (2018). “Improving same-day access in primary care: Optimal reconfiguration of appointment system setups”. In: *Operations Research for Health Care* 18, pp. 119–134. DOI: 10.1016/j.orhc.2017.09.003 (cit. on pp. 21, 22).

Schrijver, A. (2003). *Combinatorial Optimization: Polyhedra and Efficiency*. Vol. 24. Algorithms and Combinatorics. Berlin, Heidelberg, New York: Springer (cit. on p. 149).

Schwartz, J., A. Steger, and A. Weißl (2005). "Fast Algorithms for Weighted Bipartite Matching". In: *Experimental and Efficient Algorithms*. Ed. by S. E. Nikoletseas. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 476–487. DOI: 10.1007/11427186_41 (cit. on p. 118).

Schwartze, J. and K.-H. Wolf (2017). "Projekt "Rollende Arztpraxis" im Landkreis Wolfenbüttel"". In: *Management von Gesundheitsregionen II: Regionale Vernetzungsstrategien und Lösungsansätze zur Verbesserung der Gesundheitsversorgung*. Ed. by M. A. Pfannstiel, A. Focke, and H. Mehlich. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 81–92. DOI: 10.1007/978-3-658-12592-9_9 (cit. on pp. 77–79).

Shi, J., Y. Peng, and E. Erdem (2014). "Simulation analysis on patient visit efficiency of a typical VA primary care clinic with complex characteristics". In: *Simulation Modelling Practice and Theory* 47, pp. 165–181. DOI: 10.1016/j.simpat.2014.06.003 (cit. on pp. 21, 22).

Soyster, A. L. (1973). "Technical Note—Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming". In: *Operations Research* 21.5, pp. 1154–1157. DOI: 10.1287/opre.21.5.1154 (cit. on p. 97).

Stanford, D. A., P. Taylor, and I. Ziedins (July 2014). "Waiting time distributions in the accumulating priority queue". In: *Queueing Systems* 77.3, pp. 297–330. DOI: 10.1007/s11134-013-9382-6 (cit. on p. 62).

Tanimoto, S. L., A. Itai, and M. Rodeh (Oct. 1978). "Some Matching Problems for Bipartite Graphs". In: *Journal of the ACM* 25.4, pp. 517–525. DOI: 10.1145/322092.322093 (cit. on p. 149).

Tansel, B. Ç. (2011). "Discrete Center Problems". In: *Foundations of Location Analysis*. Ed. by H. A. Eiselt and V. Marianov. New York: Springer, pp. 79–106. DOI: 10.1007/978-1-4419-7572-0_5 (cit. on p. 83).

Thorsen, A. and R. G. McGarvey (2018). "Efficient frontiers in a frontier state: Viability of mobile dentistry services in rural areas". In: *European Journal of Operational Research* 268.3, pp. 1062–1076. DOI: 10.1016/j.ejor.2017.07.062 (cit. on pp. 77, 78, 83).

Tracy, M., M. Cerdá, and K. M. Keyes (2018). "Agent-Based Modeling in Public Health: Current Applications and Future Directions". In: *Annual Review of Public Health* 39.1, pp. 77–94. DOI: 10.1146/annurev-publhealth-040617-014317 (cit. on pp. 21, 25).

Tricoire, F., A. Graf, and W. J. Gutjahr (2012). "The bi-objective stochastic covering tour problem". In: *Computers & Operations Research* 39.7, pp. 1582–1592. DOI: 10.1016/j.cor.2011.09.009 (cit. on p. 82).

United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects 2019: Highlights*. ST/ESA/SER.A/423. eprint: https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf (cit. on p. 1).

United States Census Bureau (2017). *2017 National Population Projections Tables*. URL: https://census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html (visited on Oct. 19, 2019) (cit. on p. 1).

Valdes, J., R. E. Tarjan, and E. L. Lawler (1982). "The Recognition of Series Parallel Digraphs". In: *SIAM J. on Computing* 11.2, pp. 298–313. DOI: 10.1137/0211023. eprint: http://dx.doi.org/10.1137/0211023 (cit. on pp. 158, 161, 186, 189).

Wang, S., N. Liu, and G. Wan (2018). "Managing Appointment-based Services in the Presence of Walk-in Customers". In: *Management Science, Forthcoming*, pp. 1–41. DOI: 10.1287/mnsc.2018.3239 (cit. on p. 41).

Wiesche, L., M. Schacht, and B. Werners (Sept. 2017). "Strategies for interday appointment scheduling in primary care". In: *Health Care Management Science* 20.3, pp. 403–418. DOI: 10.1007/s10729-016-9361-7 (cit. on pp. 21, 22, 40, 41).

Wirth, H.-C. and J. Steffan (2001). "Reload cost problems: minimum diameter spanning tree". In: *Discrete Applied Mathematics* 113.1. Selected Papers: 12th Workshop on Graph-Theoretic Concepts in Com puter Science, pp. 73–85. DOI: 10.1016/S0166-218X(00)00392-9 (cit. on p. 152).

World Health Organization (2004). *ICD-10 : International Statistical Classification of Diseases and Related Health Problems*. English. 10th revision, 2nd. Geneva: World Health Organization (cit. on p. 59).

World Health Organization (2020a). *Primary health care – Why is primary health care important?* URL: https://www.who.int/news-room/fact-sheets/detail/primary-health-care (visited on Aug. 31, 2020) (cit. on p. 17).

World Health Organization (2019). *Primary health care on the road to universal health coverage: 2019 monitoring report: executive summary*. Tech. rep. Geneva, pp. 1–5. eprint: https://www.who.int/docs/default-source/documents/2019-uhc-report-executive-summary (cit. on p. 1).

World Health Organization (2020b). *Universal health coverage*. URL: https://www.who.int/healthsystems/universal_health_coverage/en/ (visited on Aug. 31, 2020) (cit. on p. 1).

Yücel, E., F. Salman, B. Bozkaya, and C. Gökalp (2018). "A data-driven optimization framework for routing mobile medical facilities". In: *Annals of Operations Research*, pp. 1–26. DOI: 10.1007/s10479-018-3058-x (cit. on p. 81).

Zellner, M. and S. D. Campbell (2015). "Planning for deep-rooted problems: What can we learn from aligning complex systems and wicked problems?" In: *Planning Theory & Practice* 16.4, pp. 457–478. DOI: 10.1080/14649357.2015.1084360 (cit. on p. 18).

Zhong, X., M. Williams, J. Li, S. A. Kraft, and J. S. Sleeth (2016). "Discrete-Event Simulation for Primary Care Redesign: Review and a Case Study". In: *Healthcare Analytics: From Data to Knowledge to Healthcare Improvement*. Ed. by H. Yang and E. K. Lee. Hoboken: John Wiley & Sons, Inc. Chap. 12, pp. 361–388. DOI: 10.1002/9781118919408.ch12 (cit. on pp. 17, 20–22).

Appendices

Appendices for Part II

A

A.1 Enforcement of Assumption 1

In Part II of this thesis, we have exclusively worked under Assumption 1, i.e., that the residual treatment capacities γ_k for $k \in L \cup P$ are non-negative. However, as pointed out in Section 8.1, this does not hold in general which would invalidate our results. Therefore, we have to explicitly enforce Assumption 1 in the master problem (MP) and thus also in the Benders formulation (Det-B) through the following set of constraints:

$$\sum_{v \in N(\ell)} u_v \, w_{v\ell} \leq \hat{b} \, x_\ell \qquad \forall \ell \in L \tag{A.1}$$

$$\sum_{v \in N(p)} u_v \, w_{vp} \leq \bar{b}_p \qquad \forall p \in P. \tag{A.2}$$

As we start to model patient demands as random variables and consider robust formulations of the SMMU, the definition of the residual treatment capacities for fixed first-stage decisions $\hat{w}$ and $\hat{x}$ has to be adjusted. That is, we define

$$\gamma_\ell \coloneqq \hat{b} \, \hat{x}_\ell - \max_{\eta \in \mathcal{U}_2} \sum_{v \in N(\ell)} \eta_v \, \hat{w}_{v\ell} \qquad \forall \ell \in L,$$

$$\gamma_p \coloneqq \bar{b}_p - \max_{\eta \in \mathcal{U}_2} \sum_{v \in N(p)} \eta_v \, \hat{w}_{vp} \qquad \forall p \in P.$$

To enforce Assumption 1 in (Rob-B), we thus have to consider the robust counterparts of inequalities (A.1) and (A.2) given by

$$\max_{\eta \in \mathcal{U}_2} \sum_{v \in N(\ell)} \eta_v \, w_{v\ell} \leq \hat{b} \, x_\ell \qquad \forall \ell \in L \tag{A.3}$$

$$\max_{\eta \in \mathcal{U}_2} \sum_{v \in N(p)} \eta_v \, w_{vp} \leq \bar{b}_p \qquad \forall p \in P. \tag{A.4}$$

Inequalities (A.3) and (A.4) are non-linear in general, and have to be reformulated in a linear way for each specific choice of the consideration set $\mathcal{U}_2$.

For interval scenarios, i.e., $\mathcal{U}_2 = H$, this is relatively straightforward as the unsteerable patient demands at each demand origin assume their upper bound which yields

$$\sum_{v \in N(\ell)} \tau_v \, w_{v\ell} \leq \hat{b} \, x_\ell \qquad \forall \ell \in L \tag{A.5}$$

$$\sum_{v \in N(p)} \tau_v \, w_{vp} \leq \bar{b}_p \qquad \forall p \in P. \tag{A.6}$$

For budgeted uncertainty sets, i.e., $\mathcal{U}_2 = \mathcal{U}_2^\Gamma$, things get slightly more complicated although we can essentially mimic our approach from Section 8.2. That is, we formulate

$$\max_{\eta \in \mathcal{U}_2^\Gamma} \sum_{v \in N(k)} \eta_v \, \hat{w}_{v\ell}$$

for fixed $k \in L \cup P$ and fixed $\hat{w}_{vk} \in \{0, 1\}$ for all $v \in V$ and $k \in N(v)$ via the following linear program:

$$\begin{aligned}
(\mathrm{P}^k_{\mathrm{LP}})(\hat{w}) \quad \max_{\eta} \quad & \sum_{v \in N(k)} \eta_v \, \hat{w}_{vk} \\
\text{s.t.} \quad & \eta_v \leq \tau_v && \forall v \in V \\
& -\eta_v \leq -\sigma_v && \forall v \in V \\
& \sum_{v \in V} \eta_v \leq \Gamma_2 \\
& \eta_v \geq 0 && \forall v \in V.
\end{aligned}$$

The dual problem of $(\mathrm{P}^k_{\mathrm{LP}})(\hat{w})$ with identical optimal solution value is then given by

$$\begin{aligned}
(\mathrm{D}^k_{\mathrm{LP}})(\hat{w}) \quad \min_{\varepsilon, \kappa, \rho} \quad & \sum_{v \in V} (\tau_v \varepsilon_v - \sigma_v \kappa_v) + \Gamma_2 \rho \\
\text{s.t.} \quad & \varepsilon_v - \kappa_v + \rho \geq \hat{w}_{vk} && \forall v \in V \\
& \varepsilon_v, \kappa_v, \rho \geq 0 && \forall v \in V.
\end{aligned}$$

Substituting the dual problem back into (A.3) and (A.4), we get the following linear set of constraints which enforce Assumption 1 for $(\mathrm{Rob}\Gamma\text{-B})$:

$$\sum_{v \in V} \left(\tau_v \varepsilon_v^\ell - \sigma_v \kappa_v^\ell \right) + \Gamma_2 \rho^\ell \leq \hat{b} \, x_\ell \qquad \forall \ell \in L \tag{A.7}$$

$$\sum_{v \in V} (\tau_v \varepsilon_v^p - \sigma_v \kappa_v^p) + \Gamma_2 \rho^p \leq \bar{b}_p \qquad \forall p \in P \tag{A.8}$$

$$\varepsilon_v^k - \kappa_v^k + \rho^k \geq \hat{w}_{vk} \qquad \forall v \in V, \, \forall k \in L \cup P \tag{A.9}$$

$$\varepsilon_v^k, \kappa_v^k, \rho^k \geq 0 \qquad \forall v \in V, \, \forall k \in L \cup P. \tag{A.10}$$

Last but not least, it remains to consider the enforcement of Assumption 1 as we disaggregate sessions; see Section 9.2. For the session-specific strategic planning problem, the residual treatment capacities for fixed first-stage decisions $\hat{w}$ and $\hat{x}$ are given by

$$\gamma_{\boldsymbol{\ell}} \coloneqq \hat{b}\,\hat{x}_{\boldsymbol{\ell}} - \sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{\ell})} u_{\boldsymbol{v}}\,\hat{w}_{\boldsymbol{v}\boldsymbol{\ell}} \qquad \forall \boldsymbol{\ell} \in \boldsymbol{L},$$

$$\gamma_{\boldsymbol{p}} \coloneqq \bar{b}_{\boldsymbol{p}} - \sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{p})} u_{\boldsymbol{v}}\,\hat{w}_{\boldsymbol{v}\boldsymbol{p}} \qquad \forall \boldsymbol{p} \in \boldsymbol{P}.$$

As a result, we can enforce Assumption 1 in $(\lambda\text{Det-B})$ using the constraints

$$\sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{\ell})} u_{\boldsymbol{v}}\,w_{\boldsymbol{v}\boldsymbol{\ell}} \leq \hat{b}\,x_{\boldsymbol{\ell}} \qquad \forall \boldsymbol{\ell} \in \boldsymbol{L}$$

$$\sum_{\boldsymbol{v} \in \boldsymbol{N}^u(\boldsymbol{p})} u_{\boldsymbol{v}}\,w_{\boldsymbol{v}\boldsymbol{p}} \leq \bar{b}_{\boldsymbol{p}} \qquad \forall \boldsymbol{p} \in \boldsymbol{P}.$$

A.2 Separation LP for (Det-B)

The separation problem for (Det-B) can be formulated as an LP based on the observations made in the proof of Theorem 8.13. That is, we need to decide whether for fixed first-stage decisions $\hat{x}$ and $\hat{w}$ there exists $U \subseteq V$ such that $\sum_{v \in U} d_v > \sum_{k \in N(U)} \gamma_k$. By encoding the choice of $U \subseteq V$ through the variables $o_v \in [0,1]$ for all $v \in V$ and the corresponding consideration set $N(U)$ through the variables $n_k \in [0,1]$ for all $k \in L \cup P$, we obtain the following formulation of the separation problem:

$$\begin{aligned}
(\text{Sep'}) \quad \max_{o,\,n} \quad & \sum_{v \in V} d_v\,o_v - \sum_{k \in L \cup P} \gamma_k\,n_k \\
\text{s.t.} \quad & n_k \geq o_v && \forall v \in V,\, \forall k \in N(v) \\
& o_v \in [0,1] && \forall v \in V \\
& n_k \in [0,1] && \forall k \in L \cup P.
\end{aligned}$$

Formulation (Sep') solves the dual problem to the LP-relation of the Benders subproblem $(\text{SP}_{\text{LP}})(\hat{y}, \hat{x}, \hat{w})$. If the optimal solution value to (Sep') is strictly positive, this yields a violated subset $\hat{U} \subseteq V$ and we must resolve the restricted master problem.

Taking a closer look at formulation (Sep'), it becomes evident that the separation problem for (Det-B) is trivial if we only consider unsteerable demands: When $d_v = 0$ for all $v \in V$ and $\gamma_k \geq 0$ for all $k \in L \cup P$ due to Assumption 1, the objective of (Sep') is obviously non-positive. Thus, the optimal solution value to (Sep') must be non-positive and there cannot exist a violated subset.

Note, that formulation (Sep') can be used to solve the separation problem for (RobI-B) by simply substituting the deterministic steerable demands $d_v \in \mathbb{N}$ by the worst case uncertain steerable demands $\beta_v \in \mathbb{N}$ for all $v \in V$.

A.3 Evaluation of Operation Cost

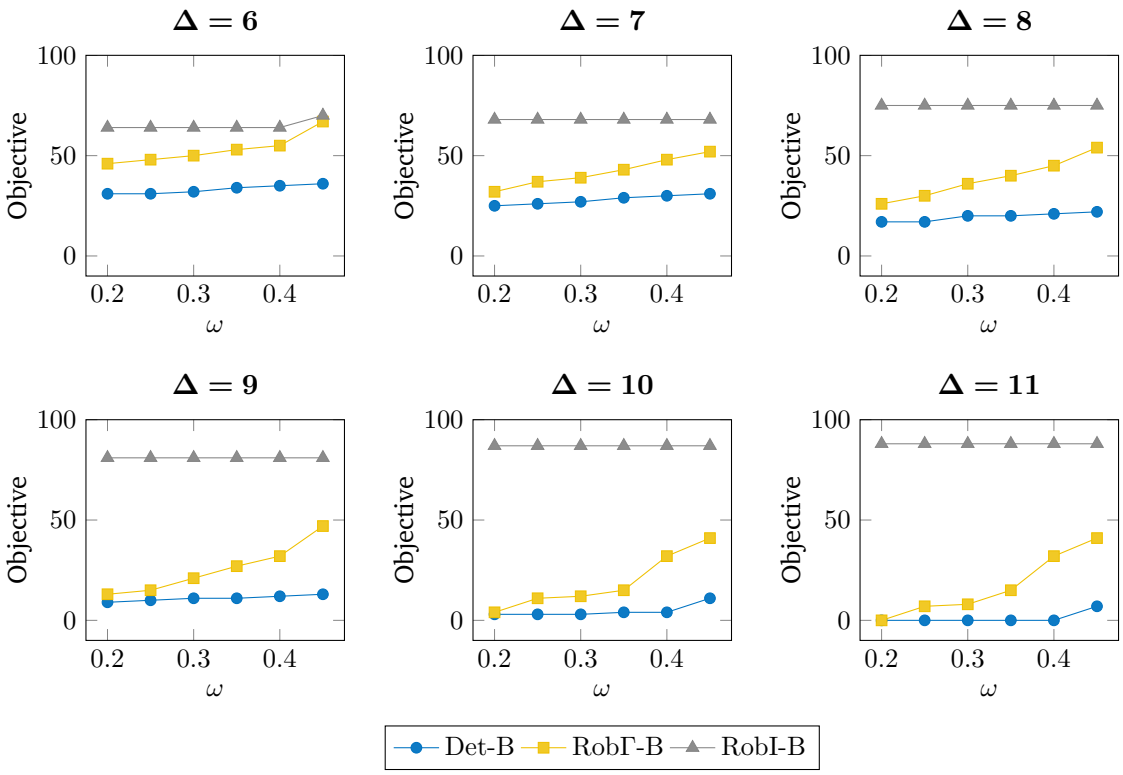


Fig. A.1.: Objectives for fixed $\Delta \in \{6, \ldots, 11\}$ and varying $\omega \in \{0.2, \ldots, 0.45\}$.

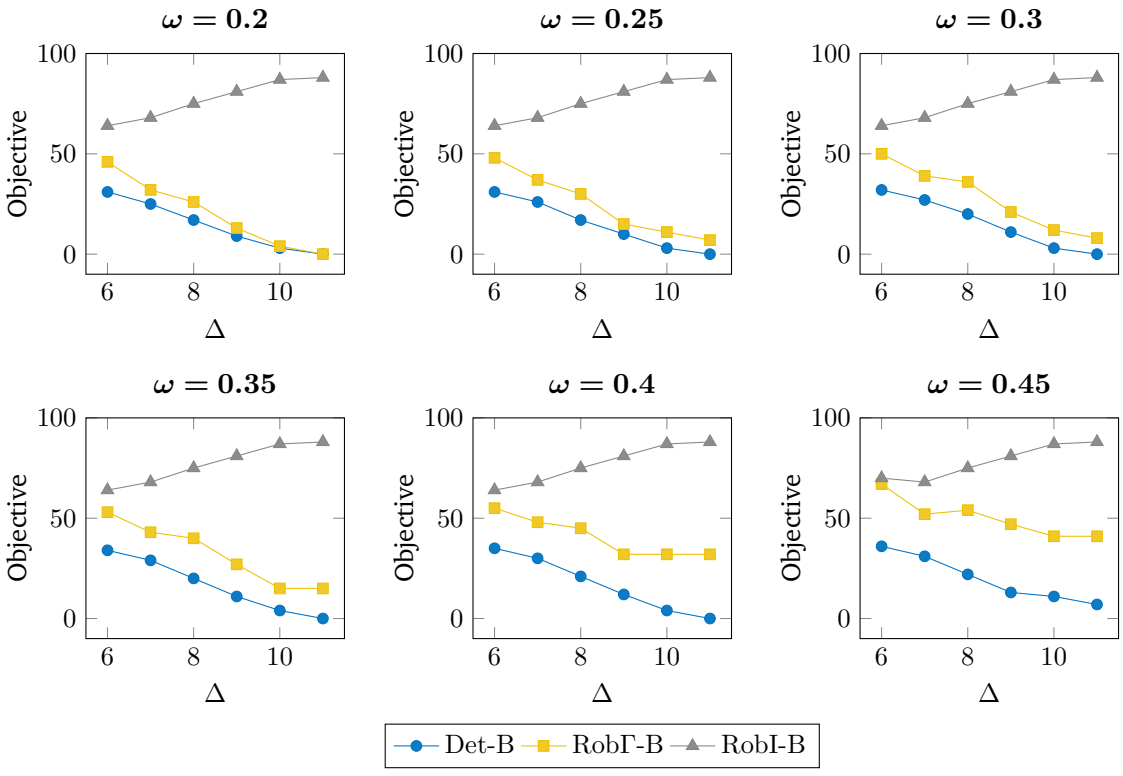


Fig. A.2.: Objectives for fixed $\omega \in \{0.2, \ldots, 0.45\}$ and varying $\Delta \in \{6, \ldots, 11\}$.

A.4 Evaluation of Solution Quality Based on SiM-Care Scenarios

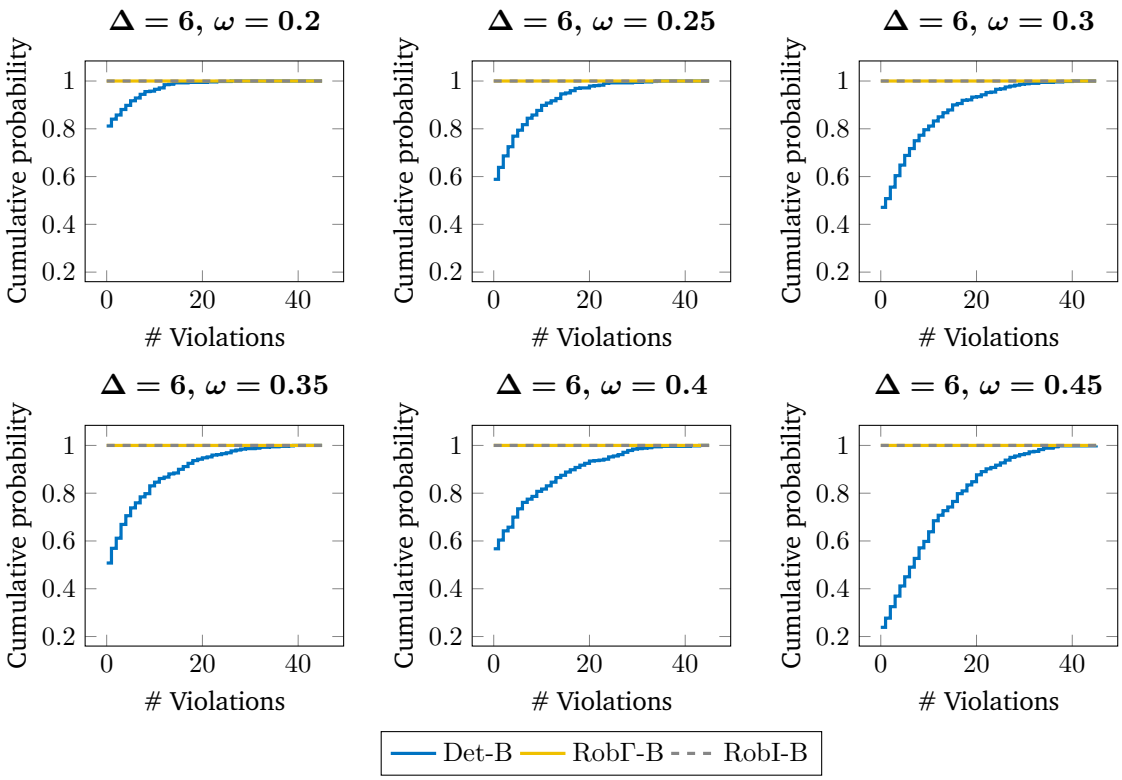


Fig. A.3.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 6$ and $\omega \in \{0.2, \ldots, 0.45\}$.

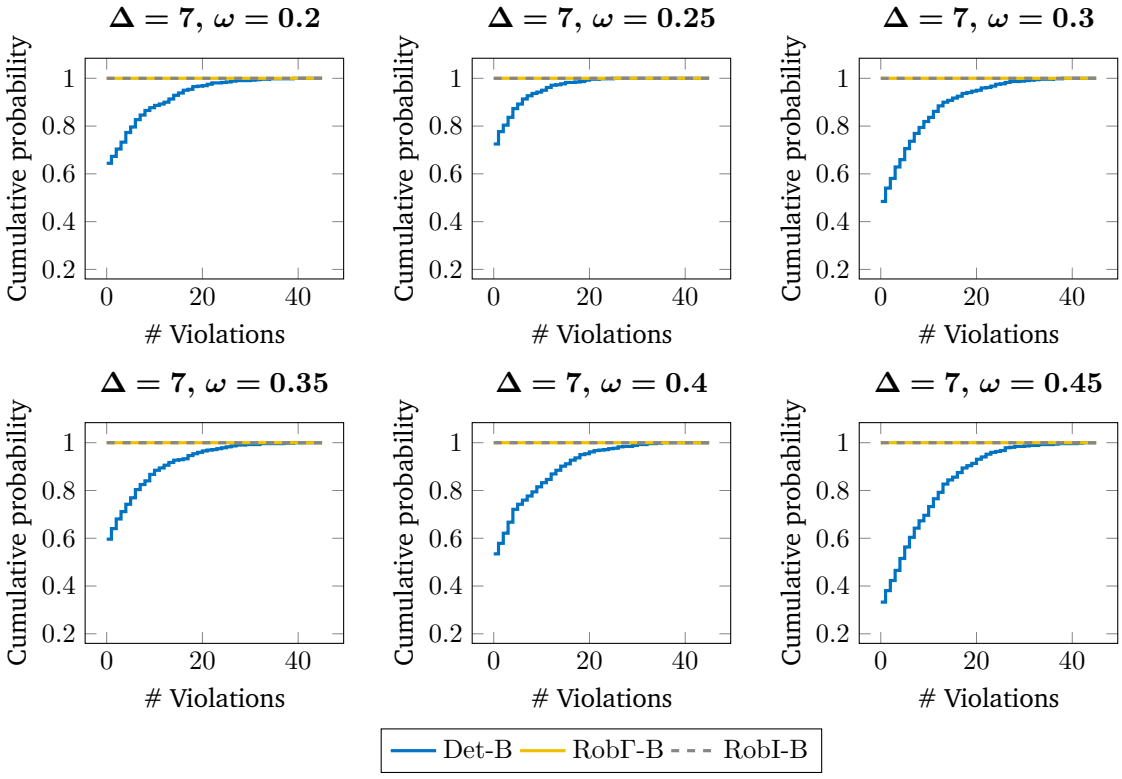


Fig. A.4.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 7$ and $\omega \in \{0.2, \ldots, 0.45\}$.

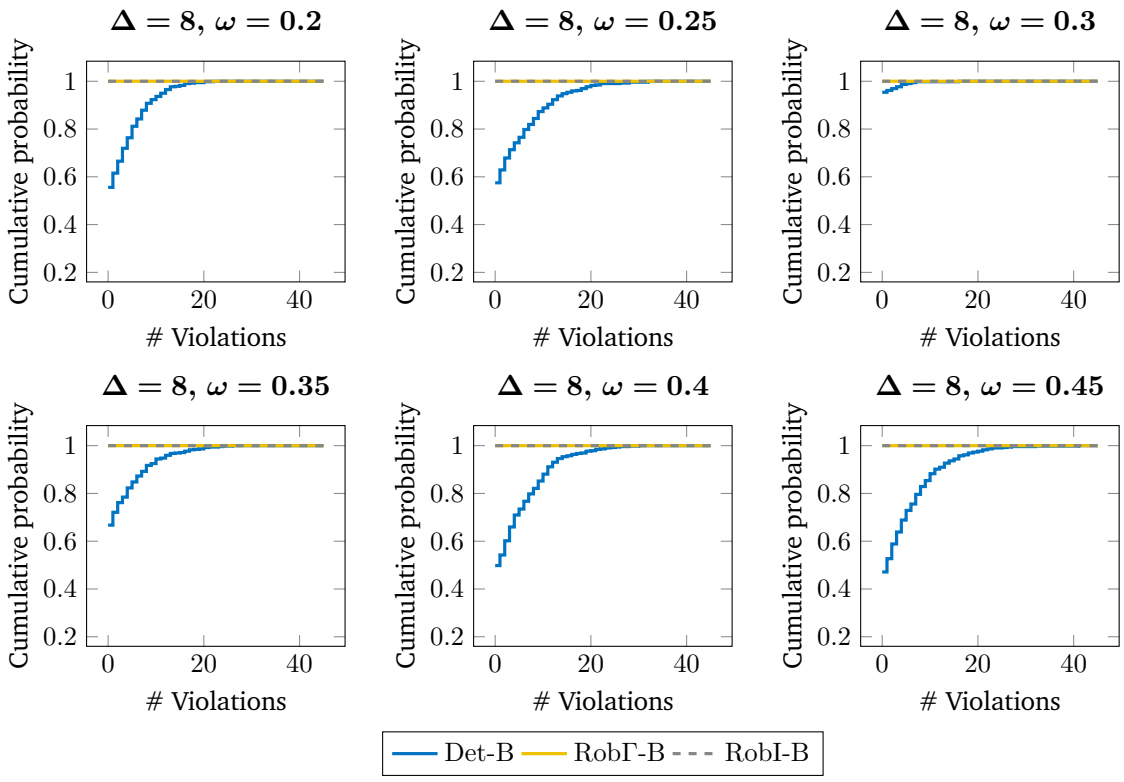


Fig. A.5.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 8$ and $\omega \in \{0.2, \ldots, 0.45\}$.

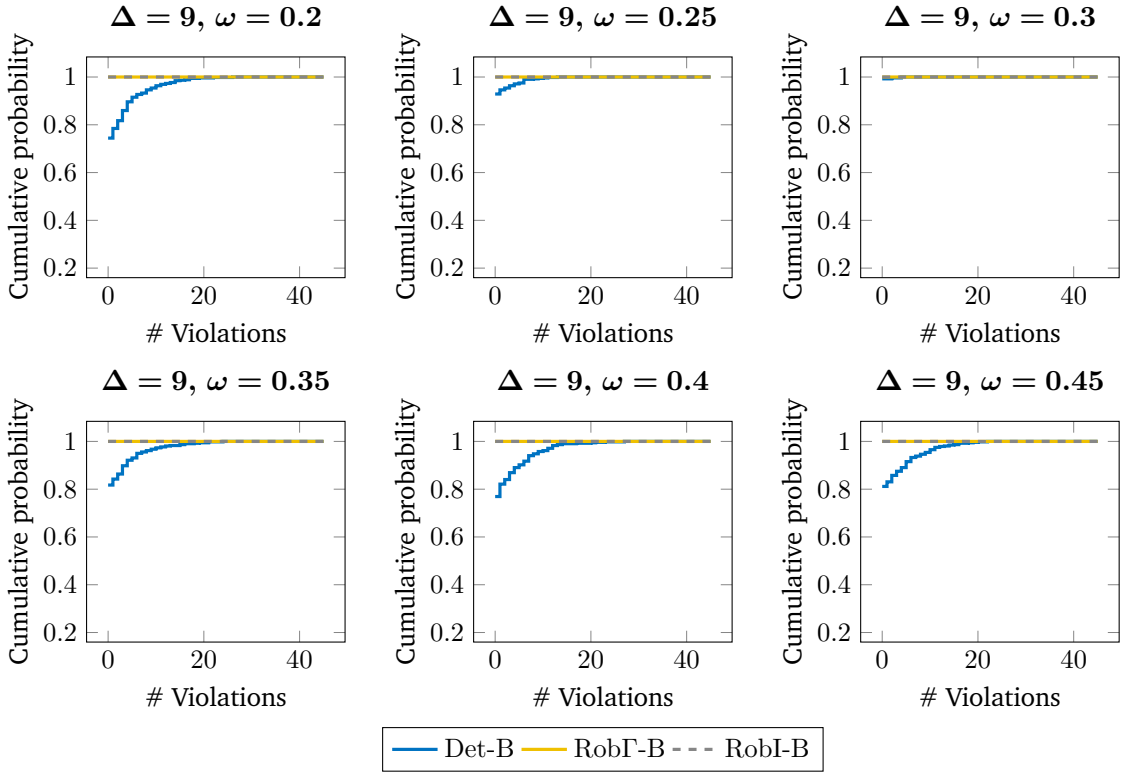


Fig. A.6.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 9$ and $\omega \in \{0.2, \ldots, 0.45\}$.

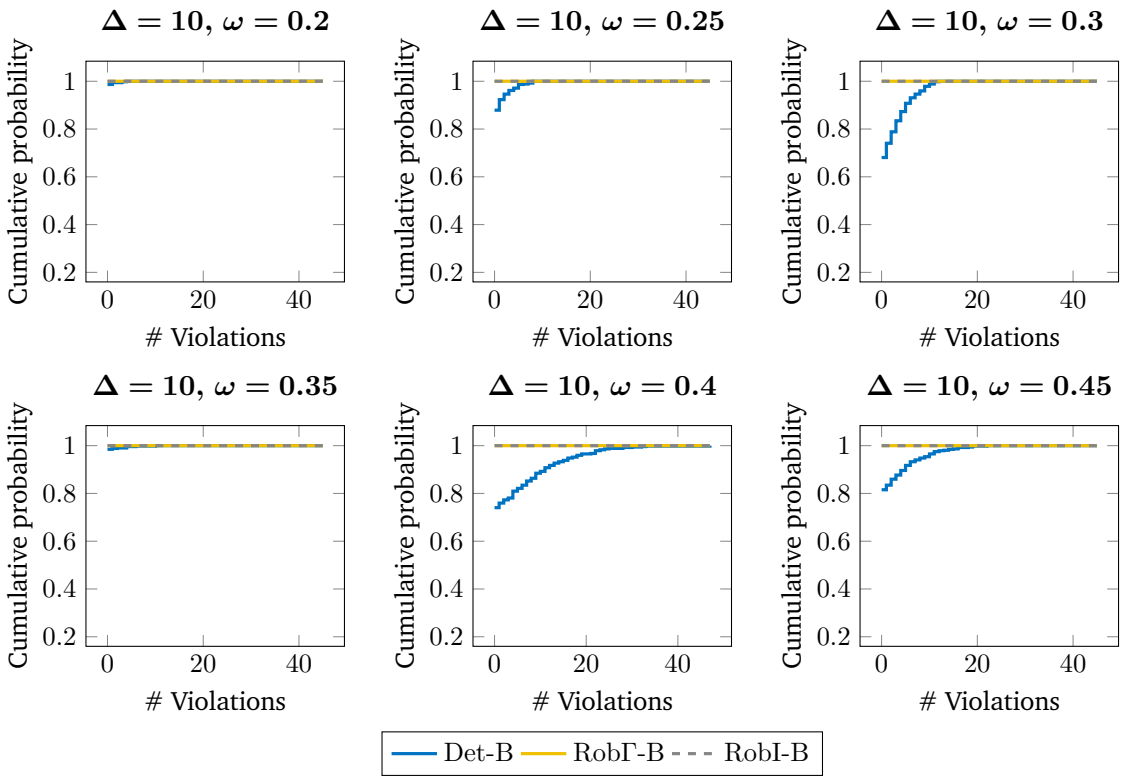


Fig. A.7.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 10$ and $\omega \in \{0.2, \ldots, 0.45\}$.

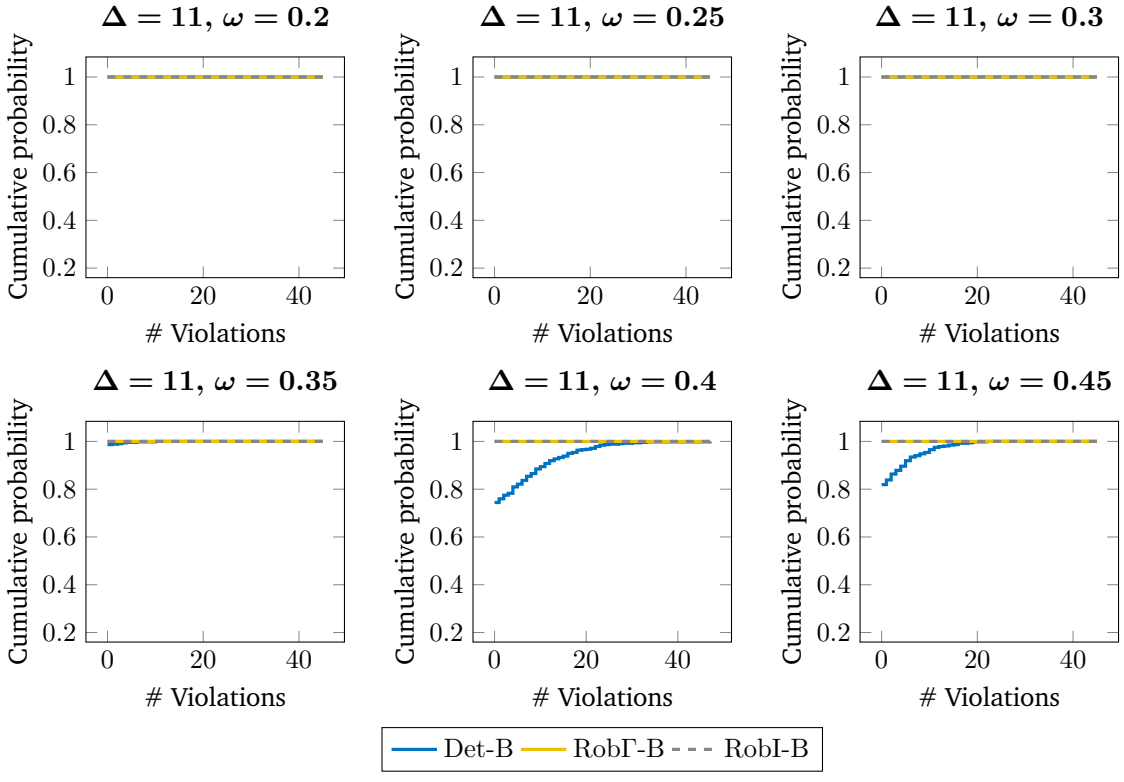


Fig. A.8.: Empirical distribution function of the minimum total number of violations for 520 realizations and parameter choices $\Delta = 11$ and $\omega \in \{0.2, \ldots, 0.45\}$.

A.5 Evaluation of Solution Quality Based on SiM-Care Scenarios with Local Surges in Demand

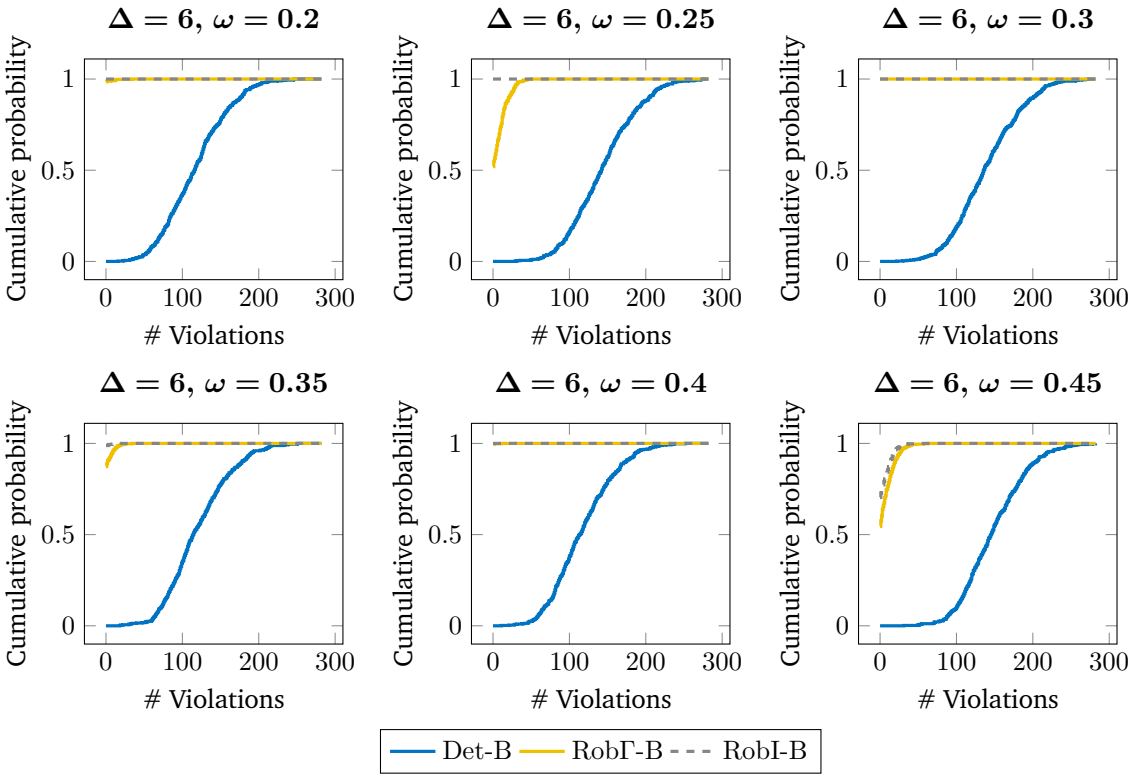


Fig. A.9.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 6$, and $\omega \in \{0.2, \ldots, 0.45\}$.

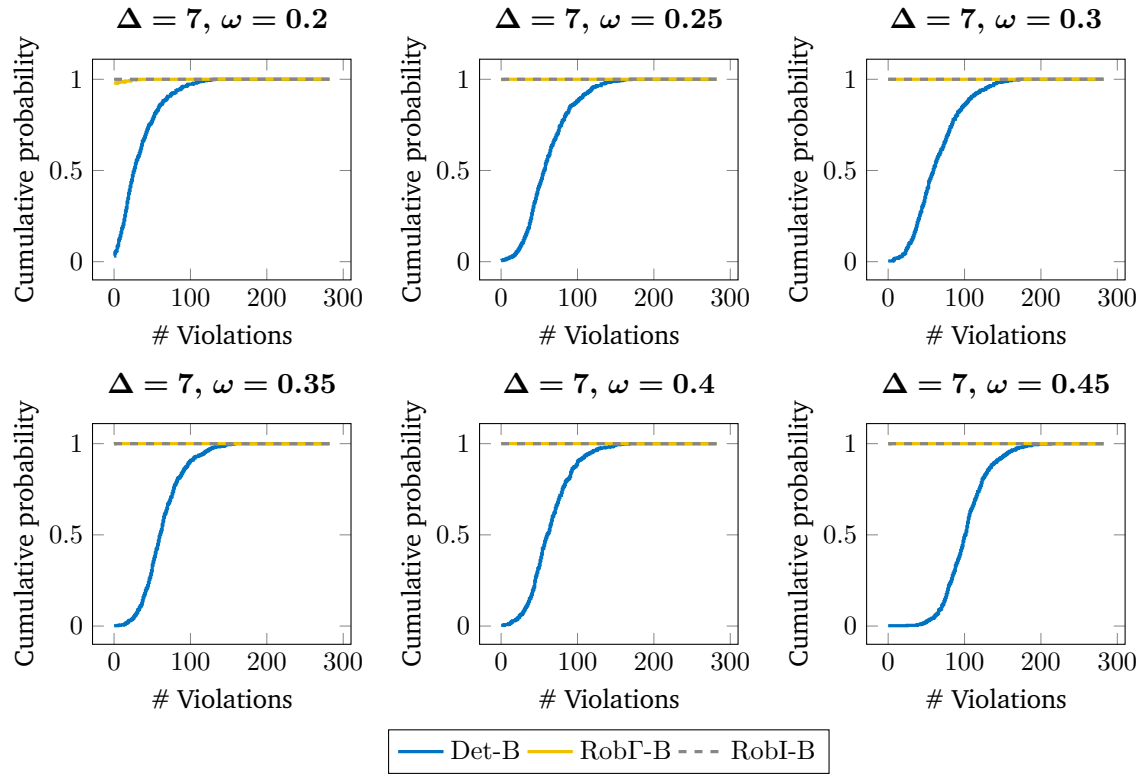


Fig. A.10.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 7$, and $\omega \in \{0.2, \ldots, 0.45\}$.

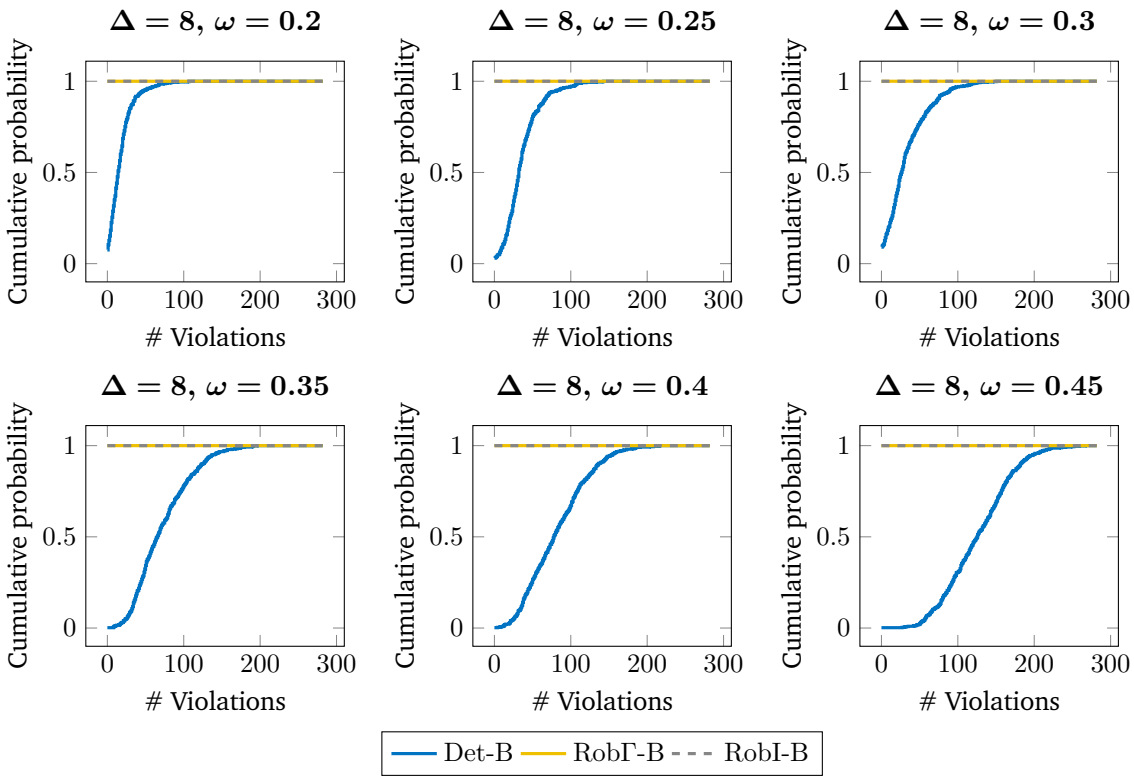


Fig. A.11.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 8$, and $\omega \in \{0.2, \ldots, 0.45\}$.

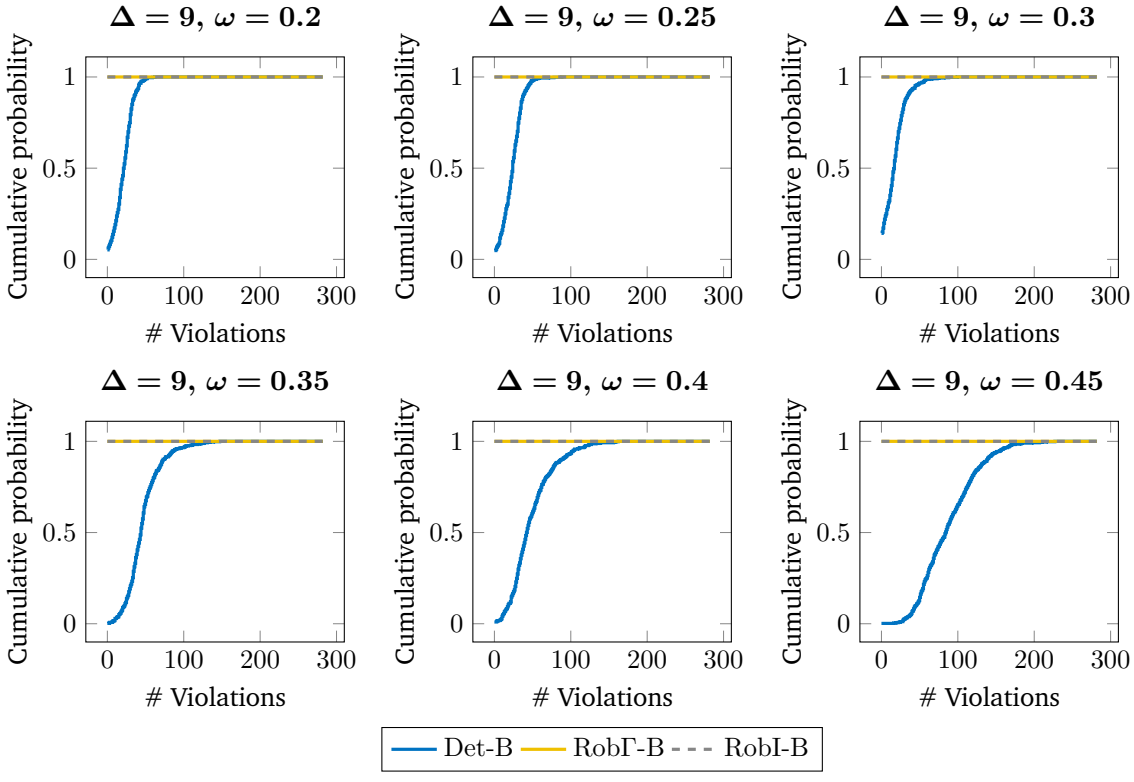


Fig. A.12.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 9$, and $\omega \in \{0.2, \ldots, 0.45\}$.

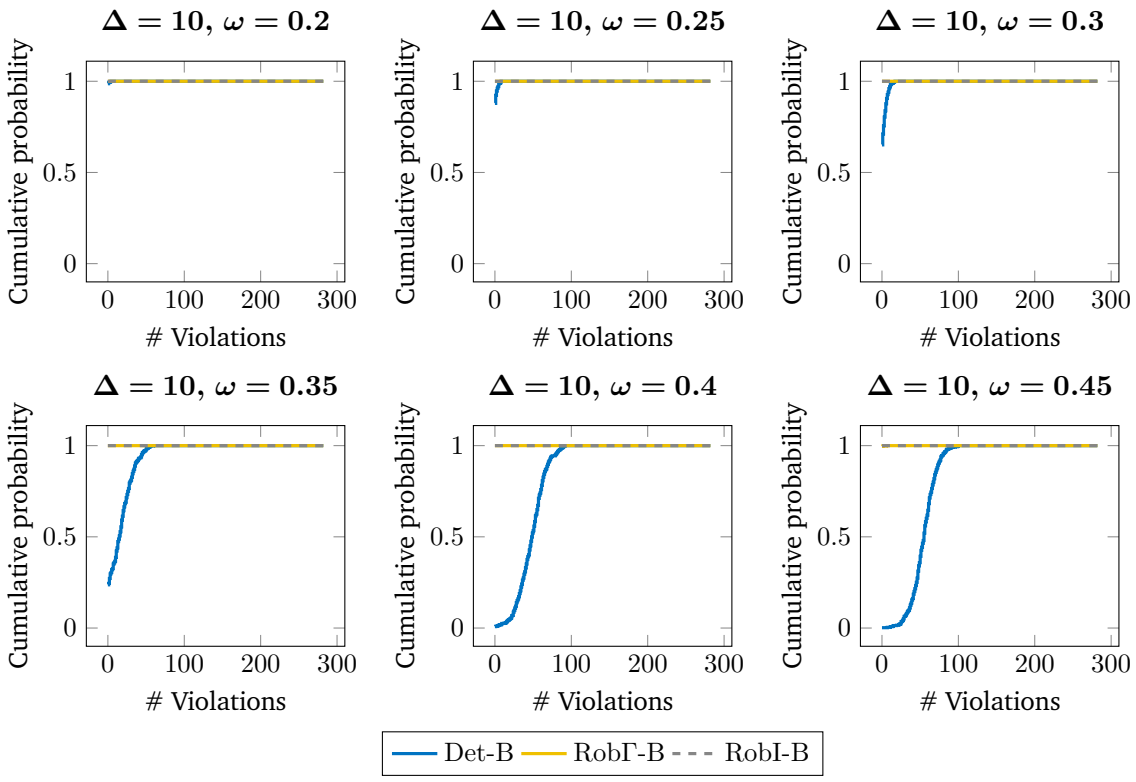


Fig. A.13.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 10$, and $\omega \in \{0.2, \ldots, 0.45\}$.

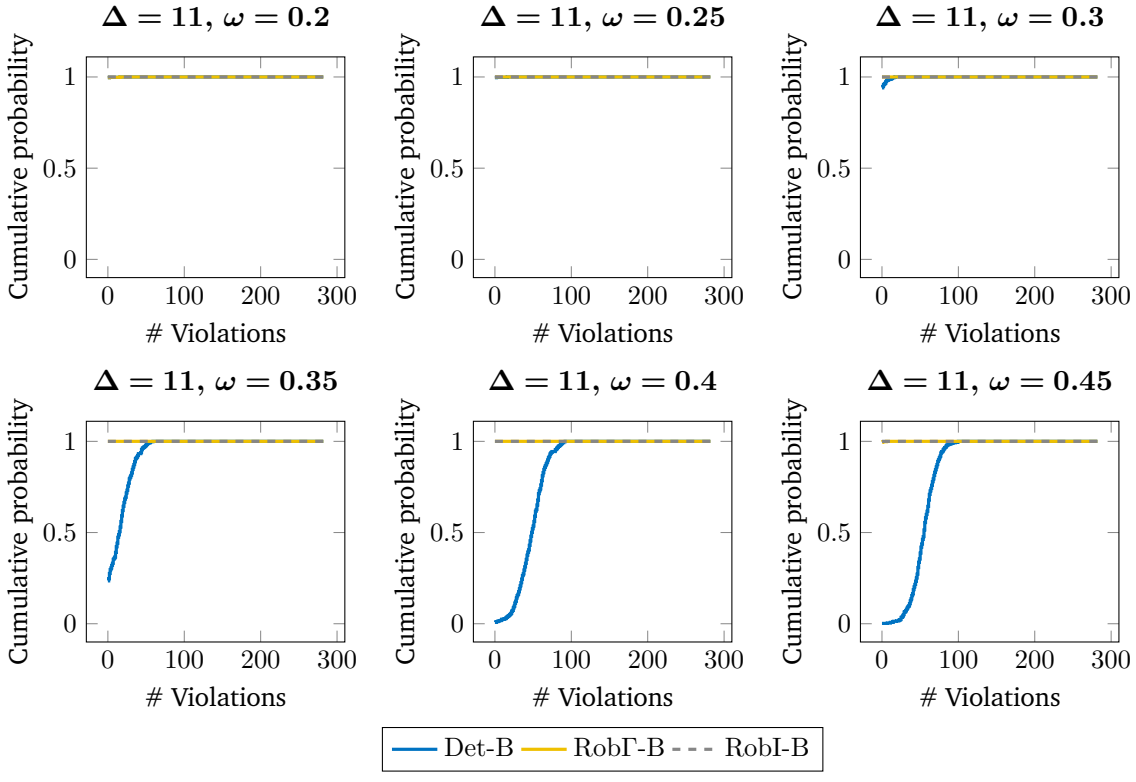


Fig. A.14.: Empirical distribution function of the minimum total number of violations for 520 realizations with 5 infectious outbreaks, $\Delta = 11$, and $\omega \in \{0.2, \ldots, 0.45\}$.

List of Figures

List of Tables

Eidesstattliche Erklärung

Ich, Martin Comis,

erkläre hiermit, dass diese Dissertation und die darin dargelegten Inhalte die eigenen sind und selbstständig, als Ergebnis der eigenen originären Forschung, generiert wurden.

Hiermit erkläre ich an Eides statt

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorand dieser Fakultät und Universität angefertigt;

2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;

3. Wenn immer andere eigene- oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;

4. Wenn aus anderen eigenen- oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Diese Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;

5. Alle wesentlichen Quellen von Unterstützung wurden benannt;

6. Wenn immer ein Teil dieser Dissertation auf der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;

7. Teile dieser Arbeit wurden zuvor veröffentlicht und zwar in:

    - M. Comis, C. Cleophas, and C. Büsing (2021). “Patients, primary care, and policy: Agent-based simulation modeling for health care decision support”. In: *Health Care Management Science*. DOI: 10.1007/s10729-021-09556-2

    - C. Büsing, M. Comis, E. Schmidt, and M. Streicher (2021). “Robust strategic planning for mobile medical units with steerable and unsteerable demands”. In: *European Journal of Operational Research* 295.1, pp. 34–50. DOI: 10.1016/j.ejor.2021.02.037

    - C. Büsing and M. Comis (2018a). “Budgeted Colored Matching Problems”. In: *Electronic Notes in Discrete Mathematics* 64. 8th International Network Optimization Conference - INOC 2017, pp. 245–254. DOI: 10.1016/j.endm.2018.01.026

- C. Büsing and M. Comis (2018b). “Multi-budgeted matching problems”. In: *Networks* 72.1, pp. 25–41. DOI: 10.1002/net.21802
- M. Anapolska, C. Büsing, and M. Comis (2018). “Minimum color-degree perfect b-matchings”. In: *16th Cologne-Twente Workshop on Graphs and Combinatorial Optimization*, pp. 13–16. eprint: http://ctw18.lipn.univ-paris13.fr/CTW18_Proceedings.pdf
- M. Anapolska, C. Büsing, M. Comis, and T. Krabs (2021). “Minimum color-degree perfect b-matchings”. In: *Networks* 77.4, pp. 477–494. DOI: 10.1002/net.21974

Martin Comis
Aachen, September 20, 2021