Maximising Achievable Throughput in Optical Network Design

A thesis submitted to UCL (University College London) for the partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

by

Robin Michael Matzner

Optical Networks Group
Department of Electronic and Electrical Engineering
UCL (University College London)

17th January 2025

Declaration

I, Robin Michael Matzner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Copyright Notice

"Die Mathematik ist es, die uns vor dem Trug
der Sinne schützt und uns den Unterschied
zwischen Schein und Wahrheit kennen lehrt"

LEONHARD EULER

Abstract

This thesis is an investigation into maximising throughput for optical core network design. To calculate the maximum achievable throughput that an optical network can sustain, an NP-hard routing optimisation problem needs to be solved and therefore designing the network to maximise this property is computationally difficult.

Both structural and physical properties impact the maximum achievable throughput of optical networks. Therefore, the SNR-BA generative graph model is proposed in Chapter 3, to investigate how structural and physical properties affect the maximum achievable throughput of optical networks. The results showed that the networks with better connectivity had on average 40% lower wavelength requirements and allocated between 8-11% more lightpaths than the SNR-BA model. However, with path lengths between 95 and 215% longer than the SNR-BA networks, they achieved 30-32% less maximum achievable throughput. This demonstrated why including physical properties within the design of optical networks is important.

More computationally efficient methods for calculating maximum achievable throughput were needed for it to be included in physical topology design. Chapter 4 explores several strategies to reduce this computational complexity, including linear programming, geometric deep learning and graph theoretical metric correlation. Results in Chapter 4 show that the proposed graph theoretical metric, demand weighted cost, had a high inverse-linear correlation to maximum achievable throughput and thus was chosen to be embedded within the optimisation problem as a proxy for maximum achievable throughput.

Chapter 5 investigates whether a proxy such as demand weighted cost can maximise the maximum achievable throughput of optical networks. Compared to a control-set the demand weighted cost showed to increase maximum achievable throughput of networks by up to 63% compared to the control-set. However, the lowest demand weighted cost did not always lead to the highest maximum achievable throughput. This showed that the objective pushes networks in the right direction, however cannot directly optimise maximum achievable throughput. To achieve this, limiting cut theory was employed, achieving a 106% increase in maximum achievable throughput compared to the control-set and thus directly optimising maximum achievable throughput of optical networks. The results of this work can be applied to future network design and to ensure intelligent access to achievable capacity.

Impact Statement

The internet has shaped society as we know it today profoundly. The infrastructure that has enabled 99% of the data transmitted world-wide is that of a web of world-wide optical networks, transmitting data via light. Traffic demand is forecasted to continue to grow exponentially and to exhaust conventional transmission media. Artificial intelligence and the extensive data requirements for its training, are a huge proportion of this traffic growth. Fundamental expansions of conventional point-to-point transmission, such as ultra-wide band and spatial division multiplexing technologies are being extensively researched currently. However, one of the fundamental limits is the physical topology of the optical network. The distribution of demands over a physical topology restricts the maximum achievable throughput. To enable future applications in artificial intelligence and beyond, requires that the design at each stage, point-to-point and network-wide, includes intelligence and therefore is optimised.

Previously, structural performance properties, such as wavelength requirements were minimised. Physical properties however affect the transmission and achievable throughput of optical networks hugely and need to be considered. Therefore maximum achievable throughput is a vital objective within designing optimal optical networks for the future.

In the research detailed in this thesis, novel generative graph modelling is used to investigate the impact of structural and physical properties on the maximum achievable throughput of optical networks. Revealed is that networks that allocated 8-11% fewer lightpaths, however with on average 58% shorter paths, achieved on average 48% higher maximum achievable throughput. This result showed that networks with good structural properties do not necessarily outperform poor physical properties. Demonstrating that these are vital to be included within the design of optical network.

The solving of the routing and wavelength assignment optimisation to calculate maximum achievable throughput of an optical network is a NP-hard problem. Therefore, integrating it into combinatorial optimisation algorithms is difficult. Several methods, including linear programming, geometric deep learning and surrogate graph theoretical metrics, therefore, were investigated to reduce the computational complexity. All avenues investigated have other future research potential. The investigated linear programming for example could enable optimal integer linear programming formulations for network design. The further development of generalisation capabilities of the geometric deep learning models trained could lead to impact in areas other than optical networking even. The final investigated graph metric, demand weighted cost was developed as a surrogate optimisation target and shown to systematically increase maximum achievable throughput by about 63%

compared to control-sets. However, this objective does not directly optimise maximum achievable throughput. Direct optimisation was shown via limiting cut theory and GN-modelling to achieve a 108% increase over the control-set.

Therefore, this work demonstrates the first systematic analysis of the role of both structural and physical properties in the maximum achievable throughput of optical core networks. This resulted in understanding how to exactly manipulate both structural and physical properties of optical networks - given a specific demand profile and distance scale - to maximise their achievable throughput. The research presented in this thesis, lays the foundations for future optimisation of optical core networks, vital for fuelling the demand of future AI workloads and other demanding applications.

Acknowledgements

The research presented throughout this thesis was not the result of my individual effort. It was a sum of effort driven by myself and supported thoroughly from a host of people that I have had the pleasure to work with over the past years. The journey of my research was not a linear one and had as many highs as it did lows.

I must thank EPSRC and the TRANSET grant, which funded my PhD, however also positioned me amongst many experienced researchers. It made me collaborate and learn hugely from a diverse set of researchers. Additionally, I must thank the Microsoft "Optics in the Cloud" programme, as another funder of my research.

I am immensely grateful to have been supervised by Polina Bayvel, who has supported and developed me into a thorough researcher, making my PhD journey a positive and worth-while experience. Thank you for supporting ideas even in their infancy and for all the last minute feedback and comments that have helped guide and improve my research hugely.

I thank all my colleagues that have helped develop novel ideas together, especially Daniel Semrau, Ruijie Luo and Henrique Buglia. It has been a pleasure to work with each and every one of you. Additionally, I would like to thank also my "social" colleagues, with whom I have shared many lunches, coffees and pints over the years.

I would like to also thank my colleagues David Saad and Yizhi Xu for many discussions on optimisation and the opportunity to work on distributed message passing optimisation. Additionally, I would like to thank Alejandra Beghelli and Georgios Zervas for many insightful conversations on networking problems and research.

Lastly I would like to thank my family for always supporting me and advising me throughout my life and especially in these last years, I know that it has not been easy. Likewise I would like to thank my partner Andrea, for always having an interest in my work, for supporting my dreams and for always giving me alternate ways of looking at life. Finally, I would like to also thank my closest of friends for keeping my spirits high and the well-deserved breaks away from UCL. This thesis would have not been possible without any of you.

Table of Contents

List of Figures

List of Tables

Acronyms

ANN artificial neural network 52

BA Barabasi-Albert 65

CO combinatorial optimisation 26

DWC demand weighted cost 127

EA evolutionary algorithm 138

EDFA erbium-doped fibre amplifiers 36

EDP edge disjoint paths 158

ER Erdos-Renyi 64

FEC forward error correction 83

FF-kSP first-fit k-shortest-paths 50

GA genetic algorithm 138

GN gaussian noise 37

GNN graph neural network 108

GPU graphics processing unit 108

GRU gated recurrent unit 113

ILP integer linear programming 27

KS Kolmogorov-Smirnoff 74

List of Symbols

Expression	Description
	Fibre loss coefficient
$_{wax}$	Distance weighting of Waxman generative graph model
$_{2}$	Group velocity dispersion
$_{3}$	Group velocity dispersion slope
$_{k}$	k-path weighting in DWC
$_{k}^{DWC}$	Path weighting constant in DWC metric
$_{wax}$	Connectivity weighting of Waxman generative graph model
$_{Katz}$	Katz index weighting factor
$_{w;k;z}$	Binary decision variable, whether a lightpath is assigned to wavelength w, over path k, between node-pair z
$_{uv}^{i}$	Whether demand with source i uses edge $(u;v)$
$_{uv}^{ij}$	Whether demand between nodes i and j uses edge $(u;v)$
	Coherence factor
$_{rl}$	Epsilon greedy factor in DQN algorithm
$_{T}$	Randomised traffic skew
	Nonlinear interference coefficient
$_{SPM}$	Nonlinear interference coefficient consisting of self-phase modulation contributions
$_{XPM}$	Nonlinear interference coefficient consisting of cross-phase modulation contributions
	Nonlinearity parameter
$_{pl}$	Power law parameter
$_{rl}$	Discount factor in Bellman equation
$_{TR}$	Distance traffic skew
	SNR_{TRX}^{1}
(G)	Set of eigenvalues of L_D resulting from a graph (G)
$!$	Edge connectivity
	Weighting of structural and physical path costs in DWC
$[h]$	Sigmoid function
$_{l}$	Distance weighting of the SNR-BA generative graph model
$_{MPNN}$	Set of parameters for MPNN

Expression	Description
	Set of parameters for target network in DQN
$(\overline{T_z^C})$	Throughput scaling factor based on the whole network
$_E(\overline{T_z^C})$	Throughput scaling factor based on edges
$_{ILP}$	Throughput scaling factor of ILP
$_N(\overline{T_z^C})$	Throughput scaling factor based on nodes
A	Adjacency matrix
A_{rl}	Set of actions for Markov decision process
$A(h)$	Matrix valued artificial neural network
a_{ij}	Adjacency matrix entry between nodes i and j, 1 if nodes are connected via a fibre and 0 otherwise
B_{CH}	Bandwidth of a transmission channel
B_{ref}	Reference bandwidth
B_R	Replay buffer
B_T	Total bandwidth of a transmission system
$b(h); i(h); j(h)$	Vector valued artificial neural networks
c	Throughput multiplier of a network
C	Capacity of a channel
C_{DWC}	Total DWC cost
$C_H^{DWC}[u;v]$	Hops DWC cost
$C_L^{DWC}[u;v]$	Physical DWC cost
C_L	Set of cuts that separate the network into two subgraphs
$C_{z;k}$	Worst-case achievable throughput between node-pair z over path k
d_h	Hidden dimension size of MPNN
d_n	Degree of node n
D	Degree matrix
D_C	Number of connection requests present
$D(G)$	Diameter of graph G
$D_l(i;j)$	Distance between i and j
D_{fibre}	Fibre distance
D_{hav}	Haversine spherical distance
D_{ij}^{LP}	Number of requested lightpaths between nodes i and j
D_{KS}	Kolmogorov-Smirnov statistical distance
D_{NET}	Total fibre distance in the network
D_s	Dispersion parameter
e_n	Eccentricity of node n
e_{sp}	Spontaneous emission factor
E	Set of edges

Expression	Description
$E_{rl}(G_t; a_t)$	Environment step function
$F(G_1; G_2; V)$	Weighted spectral distance between two graphs G_1 and G_2
f	Reference frequency
f_c	Frequency of channel c
$G(N; E)$	Graph consisting of jNj nodes and jEj edges
G_A	Amplifier gain
G_t	Graph at timestep t in Markov decision process
$GRU(h; x)$	Gated recurrent unit with state input h and input x
H	Number of servers
H_{WSD}	Set of bins used for eigenvalue estimation for weighted spectral distribution
H_k^{DWC}	Number of hops of path k
h	Planck's constant
h_n^t	Hidden vector of node n within message passing round t
I	Identity matrix
$I(e\ 2\ k_{sdk})$	Indicator function - 1 if edge e is in path k_{sdk}, 0 otherwise
$I(j\ 2\ k)$	Indicator function describing whether edge j lies on the path k, 1 if so and 0 otherwise
$I(u\ 2\ k_{sdk})$	Indicator function - 1 if node u is in path k_{sdk}, 0 otherwise
K_{sd}	Set of paths between nodes u and v
K_z	Set of paths between node-pair z
k_{sdk}	Path between source s and destination d nodes on k-th path
L	Laplacian
L_k^{DWC}	Number of spans of path k
L_D	Normalised Laplacian
L_{max}	Maximum distance present in graph
L_{NET}	Total fibre distance in the network
L_{sp}	Span length
L_{uv}	Shortest path length in terms of hops, between nodes u and v
$len(p)$	Length of path p
LP	Set of lightpaths
m	Edge addition number to every node added for Barabasi-Albert graphs
$N(n)$	Neighbourhood of node n
$M_t(h_n^t; h_u^t; e_{nu})$	Message function for message passing iteration t, constructed with hidden vectors from node n, a neighbouring node u of n and the edge feature e_{nu}
N	Set of nodes

Expression	Description
N_{CH}	Number of channels
N_{LP}	Number of lightpaths allocated
$N_{LP}^{CONTROL}$	Number of lightpaths allocated in the control-set
n_{sp}	Number of spans
p_{KS}	Kolmogorov-Smirnov test statistic
P	Launch power (also referred to as P_i, which denotes the launch power of channel i)
P_{ASE}	Amplified spontaneous emissions noise
P_{NLI}	Power originating from nonlinear interference
$Q(t)$	Complex envelope of the transmitted WDM channel
$\hat{Q}(t)$	Electric field
$Q_{rl}(s;a;\ ^{MPNN})$	Action-value function given state s, action a and parameters ^{MPNN}
$Q_{rl}(s;a)$	Optimal action value function
R	Minimum radius between nodes
$R(H_N;X_N)$	Readout function using set of node hidden vectors H_N and feature vectors X_N
R_{LP}	Number of lightpaths ratio compared to a control-set
R_P	Path length ratio compared to a control-set
R_T	Throughput ratio compared to a control-set
R_u	Number of network-facing ports for node u
R_{SL}	Number of sequential loading rounds
RWA(w)	Paths over wavelength w
SNR	Signal-to-noise ratio
SNR_{TRX}	Transceiver signal-to-noise ratio
t	Time
t_{uv}	Normalised traffic between nodes u and v
T	Maximum achievable throughput
T_{MP}	Set of message passing rounds
T_R	Raman constant
T_{rl}	Final step in epsiode
T_z	Throughput between node-pair z
$T_z^{CONTROL}$	Throughput between node-pair z of a control-set
T_z^B	Set of traffic describing the bit-rate requested between a specific node-pair z
T_z^C	Set of traffic describing the number of connection requests between a specific node-pair z

Expression	Description
$\overline{T_z^B}$	Set of normalised traffic describing the bit-rate requested between a specific node-pair z
$\overline{T_z^C}$	Set of normalised traffic describing the number of connection requests between a specific node-pair z
$U_t(h_n^t; m_n^{t+1})$	Update function for message passing iteration t using hidden vector h_n^t and the new message m_n^{t+1}
$v(G)$	Eigenvectors of L_D
$w(p)$	Cost of path p
w_{c_l}	Lower bound on wavelength requirements over cut c_l
w_o	Binary variable 1 if wavelength w is used at least once
w_r	Wavelength requirement of networks
w_{ref}	Reference wavelength
w_{SP}	Cost of shortest path between source node s and destination node d
W	Set of wavelengths
$W(G;V)$	Weighted spectrum of graph G and weighting V
$WSD(G)$	Weighted spectral distribution over set of bins H for graph G
x	Distance scaling factor
$X_u^s(\overline{T_z^C})$	Source traffic on node u resulting from traffic distribution $\overline{T_z^C}$
$X_u^t(\overline{T_z^C})$	Transit traffic on node u resulting from traffic distribution $\overline{T_z^C}$
X_E	Set of feature vectors for edge set E
X_N	Set of feature vectors for node set N
Z	Set of node-pairs in graph G

Chapter 1

Introduction

THE INTERNET has revolutionised how we live our day-to-day lives. It has enabled whole nations to stay connected during a global pandemic, allowing us to rapidly adapt to a different kind of life. It has transformed the way we research, learn and interact with each other. This societal transformation can be seen in the huge growth of demand for data, with a year-on-year growth rate of 40% between 2000 and 2018 [1]. Over the same time period broadband and mobile speeds have grown by 141% and 233%, respectively. Mobile data traffic alone is forecast to grow with a compound annual growth rate of 20% up until 2029 [2]; whilst the number of connected devices is forecast to grow by 59%, to a level of 3.6 devices per capita between 2018 and 2023 [1]. The huge growth of artificial intelligence (AI), seen with the widespread application of large language models, has compounded this traffic growth. The training of these AI models is hugely data hungry and requires parallelisation over many 100s/1000s of devices. This parallelisation involves a huge number of machine-to-machine interactions and data communication between data-centres, with a predicted compound annual growth rate of 130% between 2023 and 2030 [3]. With overall global network traffic demand growing exponentially and the distribution of demand changing with heavy data-centre to data-centre traffic skew, optical networks need to become adaptable and able to deliver capacity where and when it is needed.

This combination of increased number of devices, required data rates per device and AI, has created a phenomenon in the optical networking community often referred to as the *capacity crunch*. This is the event of future internet technologies not being able to satisfy this exponential growth in demand.

Optical fibre communications have been the key enabler in scaling this growth in internet traffic and now are the key technology in expanding future increased demand applications. Point-to-point optical communication systems have increased from 100 Mbps in 1970 to over 400 Tbps in the state-of-the-art research lab systems [5]. Key technologies to achieve this huge growth were improvements in the manufacture of optical fibres that enabled low-loss transmission windows, optical amplifiers, coherent detection, digital signal processing and wavelength division multiplexing systems.

The combination of these technologies gave way to optical fibre networks that

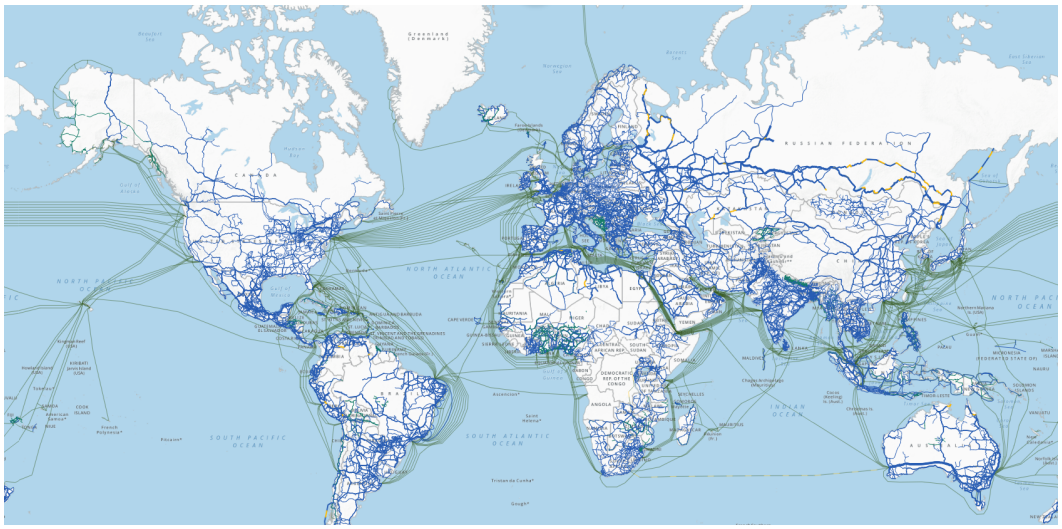


Fig. 1.1: ITU fibre infrastructure map of openly accessible optical fibre links globally [4].

transmit data on many wavelengths (100s) over many kilometres (1000s), with intermediate routing or switching nodes. Optical fibre networks have inter-connected the whole world in a web of fibre, with links spanning oceans and all continents of the world, as is seen in Figure 1.1, showing their global and universal significance. These optical networks transport the majority of internet traffic. A point-to-point lightpath within an optical network can span over many optical fibre edges and its throughput is inherently determined by its physical length in kilometres and the inter-channel distortion it experiences along its path, due to other interfering transmitted lightpaths.

In point-to-point optical communications, throughput can be increased by either exploiting larger optical bandwidths or increasing the rate of transmission of that channel [6]. These throughput gains are independent of the network topology. Although much research has investigated optimising the point-to-point transmission within optical networks, one of the limiting factors within optical networking is the physical topology, i.e. how the network nodes are physically connected with fibres.

The overall maximum achievable throughput that an optical network can sustain given a distribution of requested traffic, is limited by the structural (how the network is connected) and the physical properties (lengths/types of fibres and channel properties) of the network, and the paths and wavelengths that are assigned to lightpaths. Maximising the achievable throughput network-wide, requires both the structural and physical properties of the network to be considered.

The optimisation problems concerning how to connect the nodes and how to route demands within optical networks are referred to as the physical topology design (PTD) and the routing and wavelength assignment (RWA) problem, respectively. Both of these optimisation problems are NP-hard combinatorial optimisation (CO) problems. NP-hard combinatorial optimisation problems are a category of optimisation problems that

have no algorithms that can solve these problems exactly within polynomial-time. This makes them computationally expensive to solve exactly [7]. This complexity is of a nature where solving either of these optimisation problems exactly at scale (more than 30 nodes) is not only computationally complex, however computationally infeasible. Therefore, the goal is not only to speed up these optimisation problems, however rather to make the solving of these optimisation problems feasible for larger networks that are inevitable in the future.

Optical fibre networks connect AI clusters with data centres and between data centres, as machine-to-machine traffic places increased loads in the networks. Therefore, given the criticality of the optical network infrastructure in enabling the training of AI world-wide, it is necessary to intelligently design this infrastructure. Intelligence in the context of optimisation can be thought of as "optimisation power", or in other words: optimisation performance. Where intelligence needs to be able to adapt to varying traffic patterns and time/distance-scales. Including intelligence in PTD means designing them closer to or at the global optimum of maximum achievable throughput, whilst being adaptable to distance and traffic. The resulting network performance (maximum achievable throughput), depends on the search strategy used to find valid solutions.

Integer linear programming (ILP) formulations are the only method to date to solve either of the PTD and RWA problems exactly. ILPs use mathematical constraints to limit the solution space and to guide the optimisation process. However, they are only possible to be applied to small networks (10-15 nodes). Optical networks nowadays can however span up to 100 nodes [8].

Therefore, human-designed heuristics have been proposed for both of these problems. Heuristics are solution frameworks/algorithms that follow a rule-based procedure, often similar to how humans would solve a problem, rather than a mathematically optimal procedure [9]. These methods have been shown to not perform consistently close to the level of optimal methods such as integer linear programming formulations [10, 11, 12, 13].

The area of trial-and-error learning has inspired algorithms that learn policies to traverse the vast solution space and solve these CO problems [14, 15]. This generally nowadays is referred to as reinforcement learning and recently has been combined with deep learning resulting in deep reinforcement learning. Deep reinforcement learning has much celebrated success recently and has inspired its application to many scientific areas, optical networking not being an exception [16, 17].

One problem is how to search the solution space and find the best solution, given a large number of valid combinations. Another is how to quantify the performance of a network design in terms of maximum achievable throughput.

In the 1990s, optical networks emerged as a promising approach to optical network

operation by operating the network in the wavelength domain with wavelengths used for routing. Initially a modest number of wavelengths were used (10s) and physical properties were ignored. Therefore, PTD research focused on minimising the number of wavelengths required to allocate a given traffic, termed wavelength requirements [18, 11, 10]. Therefore, specific design methods that optimise this performance property are well-known to date. As traffic requirements increased, it became clear that the physical properties and the maximum achievable throughput of the network had to be taken into account. For maximum achievable throughput, both the structural and physical properties affect the resulting throughput and need to be taken into account in the network analysis and synthesis process. Therefore, it is required to understand what role both the structural and physical properties play within maximising the achievable throughput of optical networks, not known at the start of this PhD research.

In addition, the evaluation of maximum achievable throughput, requires the optimal evaluation of the RWA problem. This scales the computational complexity of this problem further and, therefore, has resulted in the general exclusion of this design objective [19, 20, 21]. Therefore, other computationally efficient and accurate methods for the calculation of maximum achievable throughput are required for it to be included in PTD. Machine learning, in specific, geometric deep learning has shown recent promise in learning complex graph properties, with ultra-fast inference times [22, 23, 24]. In addition to this, previous optimisation of wavelength requirements focused on optimising surrogate objectives, such as the second smallest eigenvalue of the Laplacian - termed algebraic connectivity [25, 26]. Evaluating these methods and including them within the maximisation of achievable throughput of optical networks is vital for expanding throughput of optical networks intelligently for the future.

Including intelligence in maximising the overall achievable throughput of optical networks through the design of the physical topology, means designing networks that directly optimise this property. Whether this intelligence is human (heuristic), mathematical or artificial is not important, only that it directly translates to optimal performance.

The overarching goal of this thesis is to maximise achievable throughput within optical core networks. The research first focused on investigating the roles played by the structural and physical properties of optical networks in maximum achievable throughput. In the process, a novel generative graph model is developed to model realistic optical core networks and to investigate a variety of structural and physical properties.

The next step was to investigate several computationally efficient methodologies for integrating the maximum achievable throughput of networks into the PTD problem. One of these frameworks is then investigated within several combinatorial optimisation algorithms for optimising the maximum achievable throughput of

networks. Through these simulations it is shown that it is possible to directly optimise the maximum achievable throughput of optical networks, with the goal of intelligently future-proofing this expensive global communication infrastructure to enable future communications, AI and other applications.

1.1 Thesis Outline

The rest of the thesis is structured as follows.

Chapter 2 explores the background and theory for point-to-point optical communication systems, physical layer impairments modelling, wavelength routed optical networks (WRON), the routing and wavelength assignment problem and the physical topology design problem.

Chapter 3 presents the problem of generating topologies according to real topology distributions and introduces a new generative graph model that is used to produce large synthetic datasets. This model is then used to investigate the effect of structural and physical properties on the maximum achievable throughput of optical networks.

Chapter 4 investigates how to calculate the performance of optical networks, specifically maximum achievable throughput in computationally efficient ways. Here three separate investigations are presented, each with their own strengths and weaknesses. Linear programming, geometric deep learning and analytical graph metrics are investigated as methods to include maximum achievable throughput in the PTD problem.

Chapter 5 investigates whether methods from Chapter 4 can maximise achievable throughput when used within network design algorithms. The investigation uses one of the computationally efficient objectives and researches whether direct optimisation is possible.

Finally, conclusions of this thesis and future directions are presented in *Chapter 6*.

1.2 Key Contributions

The research work described in this thesis led to the following key results:

- Derivation of a generative graph model that recreates structures and physical properties prevalent in real optical core networks. This model was tested and compared against already existing generative graph models that were previously investigated. This model allows for (i) generation of larger sets of graphs to test algorithms on (ii) investigation of the impact of optical network structures and physical properties on performance (iii) generate training data for machine learning applications. These results led to the publications in [P1, P5] and are presented in Chapter 3.

- Derivation and proof of computationally efficient optical core network design objectives equivalent to that of maximum achievable throughput. The research investigated network metrics that correlate well with maximum achievable throughput of optical networks. A metric, termed demand weighted cost (DWC) is derived for maximising the achievable throughput of optical networks. This metric is determinable in linear-time and is able to be included in traditional optimisation heuristics and meta-heuristics. This work is described in Chapter 4 and led to the publications [P2, P7].

- Integration of maximum achievable throughput in the design of optical core networks, through the development of a graph theoretical metrics, demand weighted cost. Using the demand weighted cost, the maximum achievable throughput was maximised for optical core networks. This is the first time that topology design considered this problem to the best of my knowledge. This work is described in Chapter 5.

- Using geometric deep learning it was for the first time applied to learning optical core network performance in terms of maximum achievable throughput. The benefits and limitations of using machine learning for ultra-fast performance estimation were demonstrated. This research is described in Chapter 4 and led to the publications [P3, P8, P9].

- For the training of the geometric deep learning algorithm, a large set of topologies with mostly optimal routing and wavelength assignment solutions were calculated. This dataset is invaluable for future machine learning work in the field. This work is described in Chapter 4 and used in publications [P3, P8, P9].

- Limiting cut theory along with closed-form physical layer modelling is shown to directly optimise the maximum achievable throughput of optical networks. This result is an extension of well-known results presented already in [18], however for the more complicated objective of maximum achievable throughput rather than wavelength requirements. This work is described in Chapter 5.

- Topology Bench: An open-source dataset documenting an exhaustive list of 105 real optical network topologies, and 270,000 synthetic network topologies. In addition to the open-source dataset, graph theoretical analysis of these networks, as well as guidance for future research as to how to select diverse topologies (the importance of which is demonstrated at the beginning of Chapter 3) for research, led to the publication [P4].

1.3 List of Publications

A subset of the work presented in this thesis was first published in the following academic publications:

Journal papers

P1. **Robin Matzner**, Daniel Semrau, Ruijie Luo, Georgios Zervas and Polina Bayvel, "Making intelligent topology design choices: understanding structural and physical property performance implications in optical networks [Invited]", Journal of Optical Communications and Networking, vol. 13, no. 8, pp. D53-D67, August 2021.

P2. Ruijie Luo, **Robin Matzner**, Alessandro Ottino, Georgios Zervas, and Polina Bayvel, "Exploring the relationship among traffic, topology, and throughput: towards a traffic-optimal optical network topology design", Journal of Optical Communications and Networking, vol. 15, B1-B10 (2023).

P3. **Robin Matzner**, Ruijie Luo, Georgios Zervas, Polina Bayvel; "Intelligent performance inference: A graph neural network approach to modeling maximum achievable throughput in optical networks", APL Machine Learning, 1 June 2023; 1 (2): 026112.

P4. **Robin Matzner**, Akanksha Ahuja, Rasoul Sadeghi, Michael Doherty, Seb J. Savory, and Polina Bayvel, "Topology Bench: Systematic Graph Based Benchmarking for Core Optical Networks", Journal of Optical Communications and Networking, December 2024.

Conference papers

P5. Polina Bayvel, Ruijie Luo, **Robin Matzner**, Daniel Semrau and Georgios Zervas, "[Invited] Intelligent design of optical networks: which topology features help maximise throughput in the nonlinear regime?", 2020 European Conference on Optical Communications (ECOC 2020), Brussels, Belgium, 2020, pp. 1-4.

P6. Ruijie Luo, Yi-Zhi. Xu, **Robin Matzner**, Georgios Zervas, David Saad and Polina Bayvel, "Message Passing: Towards Low-Complexity, Global Optimal Routing and Wavelength Assignment Solutions for Optical Networks", 2022 Optical Fiber Communications Conference and Exhibition (OFC 2022), San Diego, CA, USA, 2022, pp. 1-3.

P7. Ruijie Luo, **Robin Matzner**, Georgios Zervas and Polina Bayvel, "[Invited] Towards a Traffic-Optimal Large-Scale Optical Network Topology Design",

2022 International Conference on Optical Network Design and Modeling (ONDM 2022), Warsaw, Poland, 2022, pp. 1-3.

P8. **Robin Matzner**, Ruijie Luo, Georgios Zervas and Polina Bayvel, "Ultra-fast Optical Network Throughput Prediction using Graph Neural Networks", 2022 International Conference on Optical Network Design and Modeling (ONDM 2022), Warsaw, Poland, 2022, pp. 1-3.

P9. **Robin Matzner**, Ruijie Luo, Georgios Zervas and Polina Bayvel, "Expanding Graph Neural Networks for Ultra-Fast Optical Core Network Throughput Prediction to Large Node Scales", 2022 European Conference on Optical Communication (ECOC 2022), Basel, Switzerland, 2022, pp. 1-4.

P10. **Robin Matzner**, Henrique Buglia and Polina Bayvel, "Evolving Optical Core Networks: Understanding the Impact of Topology Redesign using Space and Wavelength Domains on Network Throughput", 2023 European Conference on Optical Communication (ECOC 2023), Glasgow, UK, 2023.

Chapter 2

Optical Networking Background Theory

This Chapter introduces some main theoretical concepts that are necessary to understand the rest of this thesis. It focuses on the concept of modelling optical networks as graphs and how the nodes and links are modelled and their equivalent real-world physical architectures and how data communication channels are modelled over optical fibres. In addition, the concept of wavelength routing is introduced, describing the two most difficult problems within optical networking: (i) which wavelengths and routes to choose given particular traffic demands and (ii) how to physically connect the network.

2.1 Optical Networks

Optical networks are formed by sets of point-to-point links. The transmission over these point-to-point links and the physical effects that affect the transmitted signals need to be understood. The next section covers the transmission of data over point-to-point optical fibre links.

2.1.1 Optical Point-to-Point Transmission

An optical point-to-point transmission system generally is comprised of an optical transmitter, fibre spans (n_{sp}), optical amplifiers and an optical receiver, shown in Figure 2.1. Data is modulated using amplitude, phase and polarisation of the light to transmit it over the fibre, after which it is then received and decoded at the receiver side, to then be converted back into the electrical domain.

Generally, light is modulated on a specific wavelength channel of a wavelength division multiplexed system (WDM). The electric field $\hat{Q}(t)$ of this WDM channel of frequency f_c is described by

$$\hat{Q}(t) = Q(t)e^{\ j2\ f_c t} \tag{2.1}$$

where $Q(t)$ is termed the complex envelope of the transmitted WDM channel. The evolution of this complex envelope is well-described by the nonlinear Schrödinger equation (NLSE) [27], defined in Eq.(2.2).

Here, is the loss coefficient, $_2$ and $_3$ are the group velocity dispersion and slope parameter respectively at a reference wavelength w_{ref}, denotes the nonlinearity parameter and T_R is the Raman time constant.

$$\frac{@Q}{@z} + \frac{}{2}Q + j\frac{_2}{2}\frac{@^2Q}{@T^2} \quad \frac{_3}{6}\frac{@^3Q}{@T^3} = j \quad jQj^2Q \quad T_R\frac{@jQj^2}{@T}Q \tag{2.2}$$

This can be re-written in the frequency domain by applying the Fourier transform

$$\begin{aligned}\frac{@}{@z}Q(f) = & \quad \frac{}{2} + j2 \ ^2 \ _2f^2 + j\frac{4}{3} \ ^3 \ _3f^3 \quad Q(f) \\ & + j \quad fQ(f) \sim Q(\quad f) \sim Q(f) \quad jT_R[2 \quad f(Q(f) \sim Q(\quad f))] \sim Q(f)g\end{aligned} \tag{2.3}$$

The NLSE is only valid for one polarisation however, where in communications both polarisations are normally used and therefore, this is expanded by the Manakov equation [28]. To simulate the propagation of light along the fibre one needs to solve one of these two equations.

These equations describe the evolution of a pulse through the fibre, where both linear and nonlinear effects affect this pulse. The linear effects can be grouped into attenuation and dispersion; where attenuation generally results mainly from material absorption and Rayleigh scattering and dispersion results in pulse broadening, due to different spectral components of the pulse travelling at different group velocities. These however are linear with respect to $Q(t)$. However, when light travels at high power through an optical fibre, the fibre response becomes highly nonlinear. These nonlinear effects are the result of the nonlinear Kerr effects, amongst which Raman scattering, Brouillin scattering, self-phase modulation, cross-phase modulation and four-wave mixing are the most studied.

There is no exact analytical or integral solution for neither the NLSE, nor the Manakov equation and, therefore, numerical solvers are generally used to solve these equations. This is computationally complex to simulate, although acceptable when simulating single channels over a single or multiple spans. When considering the effects on the whole network, it is necessary to simulate this propagation over many spans, channels and point-to-point links, therefore a more computationally efficient solution is required. The next section describes an approximate analytical model of NLSE, which are computationally efficient to compute. This approximate model is used to determine the signal-to-noise ratio (SNR) of lightpaths resulting from RWA solutions in Chapters 3, 4 and 5.

2.1.2 Closed-Form Physical Layer Modelling

To model the performance of data transmission over an optical fibre, it is important to take into account the noise originating from optical transceivers, optical amplifiers and

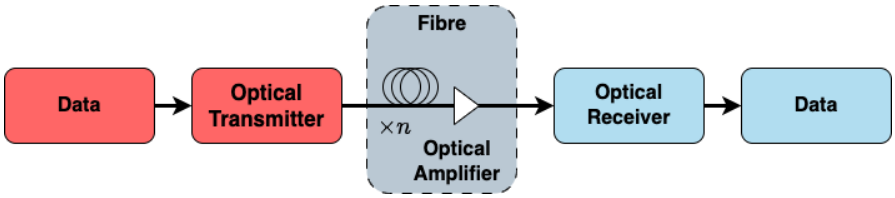


Fig. 2.1: Optical point-to-point transmission model.

fibre nonlinearities. To model the transceiver, amplifier and nonlinearity noise analytically, these three noise sources are generally modelled as independent additive Gaussian noise sources, resulting in the following definition of SNR in Eq. (2.4) [29].

$$SNR = \frac{P}{P + P_{ASE} + P_{NLI}} = \frac{P}{P + P_{ASE} + P^3} \tag{2.4}$$

Here $= SNR_{TRX}^{1}$, is the transceiver SNR, P_{ASE} is the power resulting from amplified spontaneous emission and P_{NLI} the power resulting from nonlinear interference. Generally, the transceiver noise is a significant source of noise when considering single span transmissions. When transmitting over many spans (i.e. larger distances) amplifier noise and fibre nonlinearities dominate the noise sum [30]. Therefore, as we are looking at networks that scale countries and continents, we neglect this term simplifying Eq.(2.4) to Eq.(2.5).

$$SNR = \frac{P}{P_{ASE} + P^3} \tag{2.5}$$

Using the SNR of a point-to-point transmission, one can estimate the achievable data rates over that channel, using the famous Shannon capacity formula [6].

$$C = 2B_{CH} log_2(1 + SNR) \tag{2.6}$$

Where B_{CH} is the channel bandwidth. Therefore by estimating the noise power distributions of the amplified spontaneous emission of erbium-doped fibre amplifiers (EDFA) and the nonlinearities in the fibre allows us to estimate the maximum achievable data rates. The next sections look at how one can estimate these noise profiles.

2.1.3 Amplified Spontaneous Emission

In the early days of communication over optical fibres, a signal regenerator had to be placed every 50-60 km to fully convert the optical signal back into the electrical domain, to then retransmit the data in the optical domain for the next span. This was due to the signal being attenuated in the fibre [31]. This was very costly as many transmitter-receiver pairs were necessary to communicate over long distances

(100s/1000s km). In the 1990s however, rare-earth material doped optical ampli ers, such as the EDFA were introduced, which could amplify signals in the optical domain, replacing these regenerators. In addition to the propagating signal, the EDFA ampli es spontaneous emissions which then copropogate with the signal as noise. These ampli ed spontaneous emissions are well modelled as additive white gaussian noise and its power can be calculated by

$$P_{ASE} = 2e_{sp}h(f + f_c)B_{ref}(G \quad 1) \tag{2.7}$$

where e_{sp} is the spontaneous emission factor, h is Planck's constant, f is the relative and f_c the reference frequency, B_{ref} is the reference bandwidth and G is the ampli er gain. This is a closed-form equation that can easily be included in Eq.(2.5). The next section describes how to estimate the nonlinear interference power.

2.1.4 Nonlinear Interference

There has been a large volume of work that has focused on analytically approximating the NLSE to obtain analytical solutions of it. Analytical modelling generally assumes that NLI can be modelled as Gaussian noise, and therefore is referred to as the Gaussian Noise (GN) model [32]. In particular, the GN model focuses on modelling the nonlinear coef cient from Eq.(2.4). The GN model estimates NLI by calculating the self-phase (SPM) for intra-channel distortions and cross-phase modulation (XPM) to account for inter-channel distortions. SPM accounts for the distortions imposed on a channel by itself and XPM the distortions due to cross channel interactions. Based on these assumptions, one can approximate the NLI coef cient as in Eq.(2.8) [29]. Here N_{CH} refers to the number of channels present in the transmission system and f_i is the transmission channel of interest.

$$_1(f_i) = \quad _{SPM}(f_i) + \sum_{k=1;k\neq i}^{N_{CH}} \quad ^{(k)}_{XPM}(f_i) \tag{2.8}$$

Closed-form equations have been derived for both the SPM and XPM contributions, even including inter-channel stimulated Raman scattering [29]. This effect is only necessary to include in transmissions of bandwidths exceeding the conventional band (C-band), which is approximately 5THz wide. All the work presented in this thesis focuses on transmissions over the C-band and therefore does not take into account inter-channel stimulated Raman scattering. The NLI coef cient for SPM and XPM can be approximated as in Eq.(2.9) and Eq.(2.10) respectively [29].

$$_{SPM}(f_i) \quad \frac{16}{27}\frac{^2}{B_i^2} \quad \frac{(T_i^2 \quad \frac{4}{9})}{_i}\text{asinh}\left(\frac{B_i^2 \quad _i}{16} + \frac{B_i^2}{9 \quad ^2}\right)^{\#} \tag{2.9}$$

$$ {}_{XPM}(f_i) \quad \frac{32}{27}\frac{{}^2}{} \sum_{8k2A_i} \left(\frac{P_k}{P_i}\right)^2 \frac{1}{B_k \; {}_{i;k}} \left[\frac{T_k^2 \quad 1}{3}\text{atan}\left(\frac{B_i \; {}_{i;k}}{}\right) + \frac{4 \quad T_k^2}{6}\text{atan}\left(\frac{B_i \; {}_{i;k}}{2}\right)\right] \tag{2.10}$$

By substituting Eq.(2.9) and Eq.(2.10) in Eq.(2.8), one can calculate the NLI coefficient, in presence of many WDM channels over one span, for a particular reference frequency f_i [29]. This can be extended to multi-span system by assuming that the NLI accumulates coherently as

$$ = {}_1 n_{sp}^{1+} \tag{2.11}$$

where $\in [0;1]$ is the coherence factor [33]. This closed-form GN model can be used to model the SNR characteristics of signals travelling throughout the network in presence of other channels. This means we can approximate the data rates possible over these channels. This model is an approximation to the NLSE, however, to model many channels throughout a large optical core network of 15-100 nodes, these approximations are necessary to reduce the computational complexity. This closed-form model is used to calculate the SNR and achievable throughput of lightpaths resulting from RWA solutions in Chapters 3, 4 and 5. The next section focuses on the architecture of optical metro/core networks.

Fig. 2.2: Optical metro/core network modelled as a graph.

2.1.5 Optical Metro/Core Networks

Optical core networks are the backbone for the majority of data communications. They cover distance scales from European countries (100s km) to that of whole continents (1000s/10,000s km). These are generally modelled as graphs, where nodes represent larger switching sites, as seen in Figure 2.2, from which traf c is aggregated and transmitted or received and distributed locally. The edges connecting the nodes are optical bres, which are made up of many spans. Prior to the invention of EDFAs, these spans were made up of optical bres and repeaters, to accommodate for optical attenuation within the bre [8]. Due to this attenuation, as well as the dispersion pro le over different wavelengths, and the instability of lasers at the time, WDM was not generally seen as a solution to scale capacity. As for each wavelength an additional repeater pair was required, which was expensive [31]. However, with the advent of EDFAs, more stable lasers and coherent detection and transmission, bre attenuation was overcome, accurate wavelength transmission was made possible and dispersion was mitigated through digital signal processing [34]. This enabled huge capacity scaling in the wavelength domain.

A switching architecture at the nodes was necessary to make full use of optical ampli ers and to allow non-terminating demands, i.e. transit traf c, to pass through without needing regeneration at each node. This is due to the dif culty of processing data at high frequency with electronics. Therefore, re-con gurable optical add/drop multiplexers were used to bypass electronic processing for transit traf c. The following section describes the architecture of these and explains their implication on routing within optical networks.

2.1.6 Re-con gurable Optical Add-Drop Multiplexers

At each optical core node, traf c is aggregated, which is then electronically multiplexed and modulated onto a speci c transmission wavelength. At the same time, arriving data is dropped and terminated at each of the nodes, whilst some traf c needs to pass through a node. The simplest form of node architecture maintains wavelength continuity, meaning that there is no wavelength conversion when an optical signal passes through the node. Additionally, there must be no wavelength contention on each bre, meaning that each path taken over a speci c wavelength (i.e. lightpath) on the same bre, needs to be on a unique wavelength, unused by any other lightpath. Therefore, at each node, the ROADM needs to perform three functions: (i) add connections, i.e. adding new lightpaths on to one of the incident edges (ii) drop connections, i.e. receiving lightpaths from one of the incident edges (iii) pass through, i.e. switching wavelengths from one edge to another incident edge. This can be achieved using a combination of wavelength selective switches (WSS), transmitters

and receivers, as demonstrated in Figure 2.3.

This type of ROADM architecture allows to perform wavelength routing within the network. Meaning that we can establish, terminate and pass lightpaths through nodes, without requiring optical regeneration at each node, only ampli cation. This allows for wavelength routed optical networks (WRONs), as shown in Figure 2.4. Here one can see that wavelengths w_1 and w_2 are both travelling over edge $(1, 2)$, therefore requiring two different wavelengths to avoid wavelength contention. However, the lightpath established across the path $[4, 5, 1]$ can reuse w_2 for example.

The paradigm of WRONs gave rise to a new type of operational problem, namely a combinatorial optimisation problem, referred to as the RWA problem. This problem is discussed in the next section.

2.2 Routing and Wavelength Assignment Problem

WRONs gave rise to a new type of operations research problem, i.e. the RWA problem. This problem requires the assigning of both paths and wavelengths to traf c demands.

Fig. 2.3: Route-select recon gurable-optical-add-drop-multiplexer (ROADM) architecture.

The RWA problem is known to be an NP-hard computational problem, meaning there is no algorithm to solve this problem in polynomial time [35].

There are two inputs to the RWA problem, (i) the topology (graph) $G(N, E)$ with node set N and edge set E (ii) and a traffic matrix (T_z^C and T_z^B), describing either the number of connections required between a source destination pair, or a required data rate between a source destination pair respectively. For a particular topology (G), one can calculate a set of k-shortest paths (K_z), for which there exist polynomial time algorithms such as the popular Yen's algorithm [36]. The problem can be formulated as follows:

Given a traffic matrix, with a set of connection requests T_z^C, where Z is the set of node-pairs and $z \in Z$. Then assign a path, wavelength to each request such that

$$\delta_{wkz} = \begin{cases} 1 & \text{if } (k, w) \text{ is the lightpath assignment} \\ & \text{for node pair } z \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

where δ_{wkz} is a binary variable indicating that a request between node pair z takes path k over wavelength w.

The final allocation must meet the following conditions too.

$$T_z^C = \sum_{k \in K_z} \sum_{w \in W} \delta_{wkz} \quad \forall z \in Z \tag{2.13}$$

$$\sum_{z \in Z} \sum_{k \in K_z} \delta_{wkz} I(j \in k) \leq 1 \quad \forall j \in E \quad \forall w \in W \tag{2.14}$$

Eq.(2.13) simply states that the number of connection requests are routed over a

Fig. 2.4: Example of wavelength routing, where path $[1, 2, 3]$ and $[1, 2, 4]$ require two seperate wavelengths w_1 and w_2.

correct number of discrete paths and wavelengths. Where Eq.(2.14) uses an indicator function I, which is 1 when the edge $j \in E$ lies within the path $k \in K_z$. This applies the wavelength contention constraint, meaning that no two paths can use the same wavelength over the same edges within the graph. Wavelength continuity is ensured through the direct use of paths within w_{kz}. Within this formulation a single fibre pair for each edge is assumed, however multiple fibres can be included by simply multiplying the number of wavelengths $|W(j)|$ by the number of fibres to be included. Within optical core networks without wavelength conversion this is equivalent to multiple fibre networks.

Although within the research of this thesis generally traffic in terms of lightpaths requested $T_z^{(C)}$ is assumed, bitrate requests $T_z^{(B)}$ can also easily be included. This can be done by estimating the SNR of lightpaths between a source and destination (z) pair (- with a GN-model for example - and then estimating how many lightpaths are required to satisfy a certain bitrate demand.

The difficulty of the RWA problem lies within the optimal allocation of w_{kz}, given some objective function. In the early days of optical communications, generally $T_z^{(C=1)}$ was a matrix of 1's, which demanded a singular connection between each source destination pair. The objective previously was to minimise the number of wavelengths required to route $T_z^{(C=1)}$, termed the wavelength requirement of a network. This was to reduce cost and complexity of the network, as before widespread optical amplification, each wavelength needed to be regenerated individually. Generally throughput was expanded by increasing linerates, and not by using additional wavelengths. Therefore, a wealth of research looked at minimising the wavelength requirements [18].

After the advent of optical amplification, throughput started expanding through WDM systems. The question changed from *how do we implement the network using the fewest wavelengths*, to *how do we maximise the achievable throughput with the wavelengths available*? An important distinction, as the first investigates a static set of demands, i.e. $T_z^{C=1}$, and the second investigates an unknown set of demands, i.e. the set of demands that maximises throughput within the network, given a specific traffic distribution. Therefore, it is useful to move from a matrix of integer connection demands T_z^C, to that of a demand distribution which generally can be regarded as the normalised matrix $\overline{T_z^C}$. The relationship between the two is

$$T_z^c = \lfloor \alpha \, \overline{T_z^C} \rfloor \tag{2.15}$$

where α is some scaling factor of $\overline{T_z^C}$. Finding the max scaling of this normalised matrix, is equivalent to the problem of finding the maximum set of connection requests that we can fit into our network and generally can be regarded as a throughput metric.

This problem can be regarded as what is normally termed incremental/sequential loading, i.e. connection requests are added to the network until no more t into the network (blocking occurs). This problem formulation is static in the sense that the distribution of the traf c between source-destination pairs in the network is known. However, static RWA problem formulations have full knowledge of the requests to be routed up-front, which is not the case here, where it is unknown how many requests are allocatable within the network. Currently networks are still operated at slow time-scales, where lightpaths are set up for days. However, in the future time-scales will be much shorter (seconds/minutes), where requests arrive and depart the network over the course of operation. This is known as a dynamic traf c scenario, where the exact distribution of traf c at any given time is not known. Although this differs from the assumptions within this thesis, the methodologies considered within this research are synonymous to any traf c distribution. An optical network designed to maximise achievable throughput given a speci c average traf c distribution, will still minimise blocking for a fully dynamic scenario given the same average traf c distribution, since on average it is providing the most throughput for that speci c traf c distribution.

This formulation of the RWA problem is purely structural, meaning that only the structure, i.e. how the topology (graph) is connected (via edges) demands are routed and how many demands can be routed ef ciently. This is opposed to the problem being affected by some other property of the network, such as the physical properties of the network, i.e. how long the edges are in kms for example. Physical properties add another dimension to the problem, as shorter lightpaths (in terms of physical distance) will be able to support higher data rates. In addition, due to nonlinear impairments, inter-channel interactions affect optical transmissions, therefore the congestion of the edges also affects the quality of that transmission.

In the last 10 years, bandwidth variable transmitters have matured as a technology and have enabled exible grid networks [37]. Transmitters in these systems have a ner granularity (12.5GHz versus 50GHz) and also can transmit over multiple contiguous frequency slots, depending on the requested bit-rate and the physical properties of the network. Instead of allocating paths and wavelengths, this translates to allocating paths and spectrum, referred to as the routing and spectrum assignment problem [38, 39, 40]. This problem adds another dimension to variables/constraints and therefore is a more dif cult NP-hard optimisation. Although this is a technology that is maturing now, the optimization problem of maximizing achievable throughput in optical core networks deployed with a xed grid remained a dif cult enough optimisation problem and, therefore, was focussed upon within this thesis. Fundamentally the properties that impact xed-grid networks are the same that impact exible grid networks, however the operational problems that occur are different. Therefore, all methods presented within this thesis can be expanded to include more constraints for the optimisation of exible

grid networks instead of fixed-grid networks.

2.2.1 Routing Algorithms

The problem of traversing a graph from a source node s to a destination node d is referred to as the routing problem. A path is a set of edges (or nodes), which defines the route taken. The shortest path between a source node s and a destination node d is the shortest possible path, with hops denoting the length of path in terms of number of edges traversed and physical length the distance in kilometres within that path (in optical networks). Within optical networking, multiple paths are considered, these are referred to as the k-shortest paths within the network. These multiple paths give multiple options to the algorithm doing the routing, in case a certain path is already using a specific wavelength. The following section presents algorithms for finding both the shortest paths in a graph and the k shortest paths.

2.2.1.1 Shortest Path Algorithms

A path is represented as $p = \langle n_s, ..., n_d \rangle$ for graph $G(N, E)$ with a weight function $w(p)$ defined as Eq.(2.16).

$$w(p) = \sum_{i=1}^{d} w(n_{i-1}, n_i) \tag{2.16}$$

The shortest path weight is defined as $w_{SP}(s, d)$ in Eq.(2.17).

$$w_{SP}(s, d) = \begin{cases} \min\{w(p) : s \xrightarrow{p} d\} & \text{if there is a path from } s \text{ to } d \\ 1 & \text{otherwise} \end{cases} \tag{2.17}$$

A shortest path is any path p starting at source s and ending on destination d with weight $w(p) = w_{SP}(s, d)$. The most common three shortest path algorithms are (i) Dijkstra [41] (ii) Bellman-Ford [42] and (iii) Floyd-Warshall [43]. The Bellman-Ford algorithm calculates the shortest path from the source node to all other nodes in the graph, like Dijkstra. However, the Bellman-Ford algorithm has the added benefit of incorporating negative edge weights, this however is not used within the research in this thesis. Therefore, Dijkstra runs $O(|E| + |N| \log(|N|))$, which is generally faster than the complexity of the Bellman-Ford algorithm, which $O(|N||E|)$ [44]. The Floyd-Warshall algorithm finds the shortest path between all source-destination pairs in a graph and has a runtime complexity of $O(|N|^3)$ [44]. Extending Dijkstra's algorithm to all source destination pairs in the network, requires it to be run $|N|$ times and therefore has a runtime complexity $O(|N||E| + |N|^2 \log(|N|))$, which generally performs faster [44]. For this reason, Dijkstra's algorithm is used within this thesis.

Dijkstra's algorithm works on a greedy breadth- rst search algorithm and shown in Algorithm 1. Initially the source node s is labelled with a distance $w(s;s) = 0$ (line 2 of Algorithm 1) and all other nodes $u \in N \setminus s$ are given a distance value of $w(s;u) = 1$ (line 5). All nodes are added to the queue set Q (line 7), which is then searched through. At each iteration a node v is chosen from Q (line 10), given that it has the minimum distance from the source node s. This node is then removed from set Q and its neighbourhood $N(v)$ is inspected. The distance w_s of each neighbour node u of v is calculated (line 13) and compared to the current shortest distance $w(s;u)$ (line 14). If w_s is shorter than $w(s;u)$, it replaces its value (line 15). This is repeated until the set Q is empty. This results in the shortest paths between s and all other nodes in the network.

Algorithm 1: Dijkstra's algorithm
Input : G, s
Output : p_{sd}

```
1  begin
2      w(s;s) = 0 ;
3      for u 2 N do
4          if u 6= s then
5              w(s;u) = 1 ;
6          end
7          f vg [ Q ;
8      end
9      while len(Q) 6= 0 do
10         v  argmin(w(s;n)) ;
                n2Q
11         Q  Q n v ;
12         for u 2 N (v) do
13             w_s = w(s;v) + w(v;u) ;
14             if w_s < w (s;u) then
15                 w(s;u) = w_s;
16             end
17         end
18     end
19 end
```

2.2.1.2 k-Shortest Path Algorithms

Within the RWA problem, it is necessary to know more alternate routes than just the shortest paths in the network. Shortest routes might quickly use up all wavelengths within the network and alternate routes, that do not use the same edges, might be required to route more lightpaths. Finding these routes is referred to as the k-shortest path problem. There are two widely used algorithms that are used for this purpose: (i)

Yen's algorithm [36] (ii) Eppstein's algorithm [45]. Yen's algorithm runs in $O(kjNj^3)$, whilst Eppstein's algorithm runs in $O(jEjjNj+jNj^2\log(jNj) + kjNj^2)$. Although better runtime, Eppstein's algorithm is relatively new (1997) compared to Yen's algorithm (1973) and, therefore, more widely adopted in mainstream graph theoretical code packages such as NetworkX. In addition, the computational barrier in the RWA and PTD problem mainly comes from the scaling of the combinatorial optimisation problem, rather than the path calculation. Therefore, Yen's algorithm is adopted throughout the research in this thesis to calculate k-shortest paths.

K_{sd} denotes the set of k-paths between nodes s and d, where K^k_{sd} accesses the k-th path in K_{sd}. R^k_i and S^k_i are termed the root and spur path up to node i of the k-th path K^k_{sd}. The root path is the path from source s to the node i and the spur path is the path from node i to destination node d. Another set B is used to keep track of possible paths. Yen's algorithm is demonstrated in Algorithm 2. Initially we start with the shortest path given by Dijkstra and is added to K_{sd} as the 1st k-th path (line 2 in Algorithm 2). Then each node in K^0_{sd} is iterated over (line 4) and initially the root path of R^0_i is compared with each path that exists in K_{sd}, if an overlap is found, the edge $(i; i+1)$ is removed from the graph (lines 5-7). After this, the spur path S^0_i is calculated using Dijkstra (line 9) and the new path p_{new} is generated by combining the root path R^0_i and the spur path S^0_i (line 10). If not already in B, this path is added (lines 11 and 12). After iterating over all nodes in K^0_{sd}, the path with minimum length is added to K^1_{sd} (line 17) and another $k-2$ iterations take place to find the remaining $k-1$ paths. The following sections review optimisation methods for solving the RWA problem given these k-shortest paths.

2.2.2 Integer Linear Programming

The exact solution that gives w_{kz} that maximises $\sum$ provably only exists by solving an integer linear programming formulation taking into account Eqs.(2.12), (2.13), (2.14), (2.15) and maximising $\sum_{w2W}\sum_{k2K_z}\sum_{z2Z} w_{kz}$. This is feasible for small networks, i.e. $jNj/\ 25$. However, for larger networks the worst-case computational complexity is too high [46]. This is due to the exponential scaling of the computational complexity with regards to jNj, as the node-pairs that need to be evaluated grow with the number of nodes squared ($jZj = \frac{jNj(jNj\ 1)}{2}$).

Linear programming problems were originally developed by Leonid Kantorovic. Algorithms to solve these optimisation problems optimally were then discovered by George Dantzig [47, 48]. The discrete version of linear programming is referred to as Integer linear programming formulations. These are proven to be able to solve for the global optimal solution, by using an algorithm named branch and bound, originally proposed by Land and Doig in 1960 [49]. Strictly speaking, branch and bound is a

Algorithm 2: Yen's algorithm
Input : G, s, d
Output : K_{sd}

```
1  begin
2      K_sd^0 = Dijkstra(G, s, d) ;
3      for k ∈ K do
4          for i ∈ K_sd^k do
5              for k_sd ∈ K_sd do
6                  if R_i^k ∈ k_sd then
7                      E ← E \ (i, i + 1) ;
8                  end
9                  S_i^k = Dijkstra(G, i, d) ;
10                 p_new = R_i^k ∪ S_i^k ;
11                 if p_new ∉ B then
12                     B ← {p_new} ∪ B ;
13                 end
14                 E ← E ∪ (i, i + 1) ;
15             end
16         end
17         K_sd^k ← argmin_{p_sd ∈ B} (w(p_sd)) ;
18         B ← B \ argmin_{p_sd ∈ B} (w(p_sd)) ;
19     end
20 end
```

collection of algorithms. The branch and bound algorithms are a way of dividing a difficult optimisation problem into many easier optimisation problems, forming a tree of solutions. If only branching is utilised, then the solution process resembles that of brute-force, i.e. trying each solution and seeing which is best. The reduction in computational complexity comes from the bounding element of branch and bound. Bounding makes use of the primal bounds and the dual bounds. The primal bound is the solution bound of previous nodes and the dual bound is the solution bound of the leaf node, or current explored node. If the dual bound is no better than that of the primal bound, then one can fathom that branch, meaning that one can bound this set of solutions, as no set of solutions from that node will be better than the primal. This bounding of branches, helps reduce the amount of solutions that one needs to explore, without having to compromise the optimality of a solution. How you traverse this solution tree is dependent on the type of ILP problem that one is trying to solve, however for binary search variables, which this formulation of the RWA uses, generally a depth first search is applied.

If one for example takes a line graph with three nodes, two wavelengths ($|W| = 2$) and between each of the three node pairs ($|Z_{sd}| = 3$), there is one path ($|K_z| = 1$). Figure 2.5 demonstrates the branch and bound procedure for the binary variable w_{kz}.

Here $_{wkz}$ has a length of 6, with $2^6 = 64$ possible different values that this vector could take. Generally, branch and bound is constricted as a minimisation problem, therefore, to convert the objective of $\max(\sum_z \sum_k \sum_w {}_{wkz})$ to $\min(\sum_z \sum_k \sum_w {}_{wkz})$, we use $_{wkz} = 1 \quad {}_{wkz}$, thereby inverting the objective. The incumbant is the current up-to-date feasible lower bound found. Fathoming is the process of eliminating a branch from the exploration. A branch is fathomed if the future solutions cannot be lower than the current incumbant, and a branch is impossible if no feasible future solutions exist.

Here the best search procedure is to look from right to left using a depth- rst search. This order makes sense since, if feasible, the best solution would exist on the branch furthest right ($_{wkz} = [0;0;0;0;0;0]$), remembering that this is an inverted representation of the decision variable. However, one can see that when traversing the tree along the right hand side, we come to a solution of $_{wkz} = [0;0;0;?;?;?]$, which at that point violates the wavelength contention constraint. We cannot assign paths $[1;2;3]$ and $[1;2]$ to be both on wavelength w_0. Therefore no further branching is feasible from this branch. As the tree is traversed from right to left, each impossible branch is bounded. One can see that a feasible solution is found at $_{wkz} = [0;0;1;1;0;0]$, which has an objective value of 2 (sum of $_{wkz}$). This is the new incumbant, i.e. the current feasible lower bound of the problem. One can now use this solution to bound further branches that are maybe feasible, but cannot contain a solution better than this solution. A branch bound by this criterion is fathomed.

As the tree is traversed towards the left, many branches are fathomed that might have feasible, however, not better solutions. Therefore, in total only 16 nodes are visited, instead of 64, the number required to brute-force the solution. This is maybe a modest saving, however this is the smallest instance that one can analyse for the RWA problem. Real networks are mesh networks, with 100s of node pairs, 100s of wavelengths and 10s of paths between these node pairs.

This method can provably nd an optimal solution, however essentially has a worst-case computational complexity that is equivalent to that of brute-forcing an optimal solution [46]. As one generally has a vector binary variable of $\frac{N(N-1)}{2} |K_z| |W|$, the possible solution space and therefore the branching space for branch and bound gets very large very quickly. Therefore, it is dif cult to solve this problem optimally for networks that are larger than about 30 nodes. As optical metro/core networks are growing larger and larger in size, e.g. the British-Telecom network is more than 100 nodes large or data centre networks which are often 10,000s of nodes, other methods are additionally required to scale solutions.

Fig. 2.5: Example of Branch and Bound tree traversal, which starts from the right-hand side and impossible solutions are marked in red and fathomed solutions with light blue crosses.

2.2.3 Heuristics

Heuristics are a set of "rule of thumb" algorithms which are based on domain expert knowledge to try and generalise well performing behaviours. Heuristics became a part of discussion within operations research in 1958 in Simon and Newel's paper [9]. This paper outlined their future vision for operations research, although heuristics were argued to be solutions for problems that were not well mathematically de ned. On the other hand, Minsky discussed heuristics in his 1962 "Steps toward Arti cial Intelligence" paper, for the purpose of searching and reducing computational complexity [50]. For the RWA problem to scale to larger problem instances, heuristics were developed to try and reduce the computational complexity of ILP formulations [9]. Heuristics often solve the routing and wavelength assignment problem separately. This reduces the complexity of the problem, however leads to sub-optimal results. For some optimal routing within the topology, there may not be an achievable wavelength assignment and vice versa, assuming that wavelength converters are not used within the network. In addition, the ILP is able to optimise the allocation of demands regardless of their ordering, however within heuristics generally allocations are made one-by-one and, therefore, the performance of a particular algorithm can rely heavily on their ordering. Up to this point there are no heuristics that provably give global optimal results for the RWA problem over a variety of topologies.

The simplest heuristic developed is generally referred to as k-shortest-path rst- t (kSP-FF), demonstrated in Algorithm 3 [7]. In Algorithm T^C_z is the traf c matrix, G the graph, $(s_n; d_n)$ is a tuple of source and destination requests, W is the set of wavelengths, K_z is the set of paths between source and destination z with $|K_z|$-paths, and RWA(w) returns the set of paths over wavelength w. This heuristic sequentially

Algorithm 3: kSP-FF RWA algorithm
Input: T_z^C, G
Output: RWA

```
1  begin
2      T_z^C = [(s_1; d_1); :::; (s_n; d_n)];
3      RWA = fg;
4      for z 2 Z do
5          K_z = Yen(G; z; jK_zj);
6      end
7      for z = (s; d) 2 T_z^C do
8          for k 2 K_z do
9              for w 2 W do
10                 if k 2= RWA(w) then
11                     RWA(w)  RWA(w) [ k;
12                 end
13             end
14         end
15     end
16 end
```

solves the RWA problem, by first finding a set of k-shortest paths for a source destination node-pair, generally using Yen's algorithm, an expansion of Dijkstra's algorithm [36, 41]. Both these problems are solvable in polynomial time. For each path, the lowest wavelength possible is assigned, in line with the "First-Fit" principle. If that wavelength is not assignable due to wavelength contention, one tries the next path with that same wavelength. If a path and a wavelength are found, these are assigned, otherwise one tries all the paths and keeps increasing the wavelength until an assignment is found. If no assignment is possible on any path and any wavelength, the traffic request is blocked. The kSP-FF heuristic is known to be sub-optimal compared to the ILP, however has linear computational time and scales generally with the computational complexity of the shortest path algorithm used [7].

Many algorithms have been developed to improve on the performance of this heuristic, including meta-heuristics and also machine learning algorithms such as many varieties of reinforcement learning, discussed in the next section. A simple, yet significant improvement, is possible by simply inverting the priority from the path to the wavelength within the heuristic, demonstrated in Algorithm 4. Therefore, always assigning the path which has the lowest wavelength available, thus making better use of wavelengths and packing requests more densely into the network. This heuristic is simply referred to as first-fit k-shortest paths (FF-kSP), which has been shown to perform significantly better than kSP-FF and in most cases other heuristics in general [7]. For this reason, the simulation work including heuristics to solve the RWA

Algorithm 4: FF-kSP RWA algorithm
Input : T_z^C, G
Output : RWA

```
1  begin
2      T_z^C = [(s_1; d_1); :::; (s_n; d_n)];
3      RWA = fg;
4      for z 2 Z do
5          K_z = Yen(G; z; jK_zj);
6      end
7      for z = (s; d) 2 T_z^C do
8          for w 2 W do
9              for k 2 K_z do
10                 if p 2= RWA(w) then
11                     RWA(w)  RWA(w) [ f kg;
12                 end
13             end
14         end
15     end
16 end
```

problem uses the FF-kSP heuristic. The heuristics discusses in this section use intuition from the RWA problem to nd a good solution in reasonable time. in the last decade there has been extreme interest in another approach to this problem, in speci c using machine learning algorithms with a trial-and-error approach to traversing the solution space of the RWA problem. The following section investigates this trend.

2.2.4 Reinforcement Learning

There has been a signi cant amount of work that has researched how to further improve on heuristic solutions to close the gap to optimal ILP methods [51, 52, 53, 54]. Machine learning has been proposed as a promising approach towards the goal of closing this gap. Reinforcement learning (RL), particularly, is the branch of machine learning that tackles such problems. An incredible amount of research has been into the application of RL to the RWA problem and therefore, it is presented in some depth to understand why that is and whether this is a feasible/worth-while approach for the RWA problem or other combinatorial optimisation problems within optical networking.

RL uses markov decision processes (MDP) as a way of modelling sequential decision processes. They use a state (S), action (A) and reward (R) set to model the sequential decision making process. Again, this type of method suffers from as Richard Bellman described it: "the curse of dimensionality" [55], meaning that the computational complexity scales exponentially with the size of the state variables. This can be seen with the previous formulation of the state variable s_{wkz} which scales

exponentially with node size. Therefore, ef cient ways of traversing this search space are needed.

Two learning architechtures gave way to more recent breakthroughs, namely the actor-critic and q-learning frameworks, developed by Barto, Sutton, Anderson and Watkins respectively [56, 57]. These concepts were exploited more recently, enabled through advances in computing, by combining supervised learning architecture, e.g. arti cial neural networks (ANN), with these original reinforcement learning architectures [16]. Here many agents could use ANNs within the Q-learning framework to learn policies to play many Atari games to exceptional ability.

The combination of supervised learning and reinforcement learning gave birth to deep reinforcement learning. Since this inception in 2013, there has been a urry of application of these methods to a range of optimisation problems, in hope to nd better achieving heuristics. This has been especially true in the last ve years within optical networking.

The goal of deep-RL is to learn a policy that in the long-term maximises the reward of the MDP over which it operates. The reward in the case of the RWA problem is for example the achieved blocking probability or total sum of allocated traf c. Once this policy is learnt, the algorithm can be used to solve the RWA problem with reduced computational complexity. This is because in inference, the policy (the ANN) is purely a matrix multiplication and can also be accelerated on a graphics processing units. This reduction in computational complexity compared to ILP/meta-heuristics is what is so desirable.

The original paper applying the same RL concepts to the RWA problem is often attributed to Chen's 2019 DeepRMSA paper [58], however the rst occurrence of the idea of using deep reinforcement learning for routing came about previously in 2017 by Stampa in software de ned networks and then in 2018 by Natalino in optical transport networks [59, 60]. In addition, there are multiple works that apply this idea to optical networking, appearing within months of Chen's paper. In particular, the 2019 papers by Almasan and Suarez-Varela [61, 62], both part of the Barcelona Neural Networking centre.

Since then, there has been extensive work applying different algorithms to the same end, trying to solve the RWA problem with better performance, at lower computational inference complexity [63, 64, 65, 66, 67, 17, 68, 69, 70, 71, 72]. Generally, the trend has been to show slight improvement over heuristic solutions, often the worst of heuristics, kSP-FF is used to compare performance. A recent paper has made a detailed comparison with other heuristic, meta-heuristic and ILP methods [67]. The RL agent was able to perform 13% better than the best heuristic, however still about 50% under the level of the ILP. In addition, the FF-kSP heuristic was not evaluated, which has been shown to perform better [7]. The study also shows that the

agent that learns both the routing (path assignment) and the wavelength assignment problems is not able to outperform the best-performing heuristic.

There has been a massive leap in the last 10 years, in terms of what RL can achieve. This has inspired much research into using this for solving long-standing optimisation problems. To date however, there is no real paradigm-shifting applications of RL in the routing and wavelength assignment problem. No works have shown huge gains over state-of-the-art heuristics/meta-heuristics. Some works even debate whether there is much to gain from reducing the performance gap between heuristics and optimal methods [19].

2.3 Optical Network Design Problem

In the last section, the problem of how to route lightpaths was investigated, where physical topology was an input to the problem and imposed constraints on the setup of lightpaths. The highest achievable throughput of any RWA, will always be bound by the physical topology. Therefore, the fundamental limit to any optical network performance, is the physical topology. Understanding the problem of how to design this topology for maximum performance, in terms of maximum achievable throughput, minimum latency or robustness, is important to optimise the total infrastructure.

The PTD problem, can be expressed as the following. Given a set of N nodes with physical node locations, one can define a decision variable named a_{ij}. This decision variable defines whether a node i is connected to another node j via an edge, defined in Eq.(2.18). The connectivity of a graph can be measured as in Eq.(2.19), which measures the number of edges relative to the number of possible edges in a network.

$$a_{i,j} = \begin{cases} 1 & \text{nodes } i \text{ and } j \text{ are connected via a fibre} \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

$$\rho = \frac{2|E|}{|N|(N-1)} \tag{2.19}$$

Alternatively, for optical networks a minimum connectivity requirement is that the graph is connected, i.e a single component. Additionally to this, an objective function, such as maximum achievable throughput is sought to be maximised/minimised. Therefore, the physical topology design problem culminates in finding a network topology with two specific features:

1. Feasible: satisfies connectivity constraint
2. Optimal: that no other network exists with a better pre-chosen optimisation objective.

The ideal network would be a fully connected mesh, however this is infeasible in core networks, due to deployment constraints, costs and the exponential scaling of edges ($\frac{|N|(|N|-1)}{2}$). Therefore, finding a set of edges from the fully connected network that satisfy connectivity constraints and maximise achievable throughput is computationally difficult due to the number of combinations possible.

The following section describes design objectives within optical network design.

2.3.1 Network Design Objectives

As networks evolved from electrical transmission to optical transmission and the technologies evolved for communicating data over optical fibres - so did the design objectives of the network. As technologies evolved, the design objectives of these networks generally became more complex, making the optimal criterion particularly difficult. The following section follows the evolution of network design and its objectives.

The problem of designing networks, started with the advent of centralised computing services. These communication network designs were prior to the widespread use of optical fibres for communications. Initially, purely tree-like topologies were designed and deemed best for centralised computation access [73].

However, it was then realised in the late 1960s that one could interconnect these tree-based topologies to give resource-sharing networks, i.e. mesh networks. Mesh networks offered distributed computing services, between many nodes. In mesh networks, traffic can originate in any two nodes and therefore the design problem became more complex. In these early days of networking, the main concern was latency and networks were generally designed with that in mind [74].

Throughput was also of concern, however the throughput in these networks was different to the one that we analyse in optical core networks. Throughput then depended on network flows and therefore a path's maximum throughput only depended on the edge with smallest capacity on that path. In modern optical core networks, this is not true. The throughput depends on the length of the path and the congestion along the path, due to physical layer impairments and the lack of regeneration.

With the advent of low-loss optical fibres, optical fibre communications expanded massively into computing networks. These networks still had to be regenerated every 50-60 kms, however datarates were much higher. In 1975, the most advanced coaxial cable system could transmit 274 Mbps versus the 1.7 Gbps of an optical system with repeaters every 50km, achieved in 1987 [34]. At this point it was very expensive to communicate many wavelengths, as one needed a regeneration device per wavelength [31]. Therefore, the network design objective changed to that of minimising the number of wavelengths required to connect all node pairs, i.e. wavelength requirements [18].

With this change, it became difficult to achieve networks with wavelength requirements as the optimisation objective. This is because it is much more complex to measure the wavelength requirement of a network, than maximum flow/worst-case latency. This can be seen from formulating the problem of calculating wavelength requirements. We can again use a binary variable to decide whether we use wavelength w on path k between node-pair z (λ_{wkz}, the same as in Eq.(2.12)). One can constrain the problem to only have a connection running between each node-pair, with Eq.(2.20). The same RWA wavelength contention constraints apply from Eq.(2.14).

$$\sum_{w \in W} \sum_{k \in K_z} \lambda_{wkz} = 1 \quad \forall z \in Z \tag{2.20}$$

A new variable to measure whether a wavelength has been used is introduced, namely w_o, as in Eq.(2.21).

$$w_o = \begin{cases} 1 & \text{if } \sum_{z} \sum_{k} \lambda_{wkz} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

The objective to minimise the wavelength requirement, can be defined as $\min(\sum_{w \in W} w_o)$, with the wavelength requirement, being defined as $\sum_{w \in W} w_o$. As we saw in the previous section, this optimisation problem can be exactly solved by an integer linear programming formulation, however is computationally infeasible for larger networks. This optimisation problem is a scaled down version of the maximisation problem introduced in the previous section, however to exactly calculate it for many different networks is still difficult due to the computational complexity associated with the ILP formulation.

As EDFAs became readily available, this changed the networking focus once more [34]. With EDFAs there was no need to regenerate every 50-60 kms and longer lightpaths could be established. In addition, it became a lot more cost effective to transmit over a larger number of wavelengths, as one did not need to regenerate each wavelength individually. At this point physical layer impairments became the limiting factor in network transmission. The design focus changed from minimising the number of wavelengths to that of maximising physical throughput, i.e. the maximum achievable amount of data that you can transmit throughout the network, given some specific demand and topology. This initially was investigated from an operational point-of-view in [75, 76]. However, the operational problem is limited by the physical topology with which it starts. Therefore, to fully optimise the network, this objective is required to be included in the PTD optimisation to start with.

This problem culminates in solving the RWA problem defined in the previous section by maximising $\sum_{w} \sum_{k} \sum_{z} \lambda_{wkz}$. Once the RWA is calculated, the SNR of

Fig. 2.6: Implication of edge choices on throughput, where green edge addition increases maximum achievable throughput by 14.6%, whilst the red choice only by 0.3%.

each of the assigned lightpaths needs to be calculated. This has to be done by modelling the propagation of the light for each of the lightpaths.

Two different edge additions are simulated in the NSFNET, to illustrate the importance of edge choices on the maximum achievable throughput of a network. The maximum achievable throughput of the NSFNET for two different edge choices is calculated by solving the RWA problem using the ILP formulation from section 2.2.2 and calculating the resulting SNR of all allocated lightpaths using the closed-form GN model presented in 2.1.2 using uniform traf c and shown in Figure 2.6. Its clear to see that the green edge is a much better edge to choose, due to a 14.6% increase in maximum achievable throughput, compared to just 0.4% of the red edge. This difference in maximum achievable throughput is because of the 12% increase in allocatable lightpaths when the green edge is added. When adding the red edge, the number of allocatable lightpaths stays the same, however the average path length is reduced and therefore, a slight throughput gain is achieved due to improved physical properties. An ILP was run two times in parallel, taking several hours to evaluate both of these edge choices. This demonstrates the dif culty of including this metric in the optimisation of the physical topology.

2.4 Summary

This Chapter has presented some of the basic theory necessary to understand the optimisation problems that formed the focus of the research work described in this thesis. Namely the modelling of transmission of light over a point-to-point bre connection and how to do this computationally ef ciently using closed-form expressions, instead of numerical evaluation of the NLSE or the Manakov equation. Node architecture and the resultant networks, termed wavelength routed optical networks (WRONs), were introduced. Leading into the routing and wavelength assignment problem and some theory behind solving this combinatorial optimisation problem using integer linear programming, basic heuristics and more recently deep reinforcement learning. Finally, the problem of designing the physical topology was introduced and a discussion on design objectives and the dif culty of this combinatorial optimisation problem was presented.

The design objective which is central to the research described in this thesis, is the maximum achievable throughput of a given network. Most papers investigating resource allocation schemes and the evaluation of maximum achievable throughput and topology performance, generally use a single or few real topologies, often the NSFNET [7]. The overarching goal of network design is to understand which key network topology characteristics impact the performance of the network, to then be able to change these. Therefore, to investigate and understand this question, networks of varying structure and physical properties require investigating. The next section investigates generative graph models to generate these different types of networks.

Chapter 3

Generative Graph Models for Optical Networks

The structure of optical networks, i.e. how we connect the nodes within the graph, has an impact on the outcome of the analysis on that network. Often, resource allocation studies in optical networking, use a single or a handful of real optical networks to obtain results [58, 77, 78, 79, 80, 81, 82, 83]. Amongst these, older networks often are present, such as the NSFNET. This restricts the number of topology samples that algorithms or resource allocation schemes are subjected to, diminishing the signi cance of results, i.e. are the results due to topological artefacts or an algorithmic advance?

The importance of this can be illustrated with a simple simulation. Take two graphs, G_1 and G_2, as shown in Figure 3.1. The number of lightpaths were maximised via two methods given uniform traf c and 312 available wavelengths: (i) an ILP formulation and (ii) the FF-kSP heuristic and the resulting occupation for both ILP and FF-kSP RWAs for both graphs G_1 and G_2 plotted in Figure 3.2. Here the x-axis represents the channel number, and the y-axis represents the edge (link) number. The red bars represent whether the channel (wavelength) is occupied on that edge. For graph G_1, shown in Figure 3.1(a) and its RWA occupation in Figure 3.2(a) and (b), FF-kSP was able to allocate 1828 lightpaths and the ILP achieved 1833 lightpaths. FF-kSP managed to allocate 99% of the number of lightpaths compared to the ILP for G_1. However, for graph G_2, shown in Figure 3.1(b) and its RWA occupation shown in Figure 3.2(c) and (d), FF-kSP managed to allocate 2139 lightpaths, whilst the ILP achieved 3210. Here, FF-kSP managed only to allocate 66% of the lightpaths compared to the ILP. Therefore, if the performance of FF-kSP was only measured

Fig. 3.1: Two graphs (G_1 and G_2).

Fig. 3.2: Wavelength allocations of RWAs for (a) G_1 with FF-kSP routing (b) G_1 with ILP routing (c) G_2 with FF-kSP Routing (d) G_2 with ILP routing.

based on G_1 it would perform incredibly well, whilst if only measured based on G_2 it would perform incredibly bad. This showcases that it is necessary to have a dataset of graphs to test the effectiveness of algorithms, to give results statistical signi cance.

The other motivation behind investigating generative graph models for optical networking is to understand (i) mathematical mechanisms behind the formation of structures similar to real networks and (ii) to investigate the impact of structure and physical properties of optical networks on their performance, i.e. maximum achievable throughput. For (i) it is necessary to collect a set of real networks to then compare them to generative graph models and to model their physical and structural properties. On the other hand, for (ii) its important to look at a variety of distinct structures and physical properties using multiple models, with the aim to investigate the impact of network structure and physical properties on the maximum achievable throughput.

Therefore, the following section rst introduces key graph characteristics, namely, looking at degree, diameter and spectral properties of graphs. Then existing generative graph models that have been used previously are introduced and their limitations discussed. Following this, the signal-to-noise ratio Barabasi-Albert (SNR-BA) model - developed in the course of the work described in this PhD thesis - is introduced. This model is motivated by distance dependent penalties that affect real core networks.

These graph models are compared and the results show that the SNR-BA model is closest to that of real optical core networks in terms of structure. Finally, this model is used to evaluate the impact of structure and physical properties on the maximum achievable throughput of optical networks.

3.1 Network Properties

To compare different generative graph models, one needs to be able to measure distinct characteristics of graphs. For a graph G with N nodes and E edges, one can define the graph adjacency matrix A with entries a_{ij} as below.

$$a_{ij} = \begin{cases} 1 & \text{if nodes } (i, j) \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

As this matrix defines the inherent connectivity of the graph, it can be used to formulate distinct characteristics of graph structure, such as the degree distribution, presented in the following section.

3.1.1 Degree Distributions

Each node in a graph has a specific number of edges connected to it, which is referred to as the degree d_i. This degree can be calculated via the adjacency matrix $A(N \times N)$ by summing each row, as in Eq.(3.2).

$$d_i = \sum_{j \in N_i} a_{ij} \tag{3.2}$$

Various generative graph models give different types of degree distributions, from normal to powerlaw distributions. This difference in degree distributions between models, has been a motivation for investigating generative mechanisms in graphs. This was because purely random graphs have degree distributions that are not representative of those in real networks. Therefore, real networks are often modelled to have degree distributions that are scale-free (many low degree and very small number of very high degree nodes). The intuition was mainly that many real networks (social, citation and internet networks) have degree distributions that follow a powerlaw [84]. For optical networks low-degree nodes would be a degree of two and high degree nodes of larger than 5. This however is not necessarily true for optical networks, as high degree nodes are generally very expensive.

3.1.2 Diameter Distributions

The structure of a graph also determines the paths that are available between each node pair. A path is a sequence of nodes that are traversed between a source node and destination node. A simple path metric often used, is the diameter of a graph, de ned as the longest shortest path between all node pairs in the graph. If the eccentricity (e(n)) of a noden is de ned as the maximum distance from itself to any other node within the graph, then the diameter can be de ned as in Eq. (3.3).

$$D(G) = \max_{n2N}(e(n)) \tag{3.3}$$

Well connected graphs generally will have lower diameters, whereas graphs that are sparser or are locally connected like grid graphs, will have larger diameters, in terms of hops.

3.1.3 Spectral Properties

By using a matrix representation of the graph structure, such as the adjacency matrix, one can apply linear algebra to this matrix to analyse the properties of it. This is referred to as spectral graph theory, where one analyses the structure of a graph, by looking at properties of the Laplacian of the adjacency matrix [85]. The spectra of a graph refers to the distribution of eigenvalues of either the Laplacian or its normalised version of a graph. Even though two graphs that are co-spectral (share the same spectrum) are not isomorphic, the graph spectrum has been shown to be a method for measuring similarity between graph structures [86].

The Laplacian itself is a matrix representation of the graph, like the adjacency matrix, where edges between nodes are denoted by values of -1 and the degree of nodes is noted in the diagonal of the matrix. This can be related to the adjacency matrix using Eq.(3.4) and (3.5).

$$D^{i;j} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

$$L = D \quad A \tag{3.5}$$

Whilst the simple Laplacian and adjacency matrix mainly describe the structures of regular graphs, the normalised form of the Laplacian, and its eigenvalues, allows to generalise the network properties [85]. The normalised Laplacian is de ned as in Eq.(3.6), whereI is the identity matrix.

$$L_D = I \quad D^{\frac{1}{2}}AD^{\frac{1}{2}} \tag{3.6}$$

L_D is a real symmetric matrix with real eigenvalues $\lambda(G) = \{\lambda_0, \lambda_1, \ldots, \lambda_{|N|}\}$ and real eigenvectors $v(G) = \{v_0, v_1, \ldots, v_{|N|}\}$. These eigenvalues lie in the range 0 to 2. The distribution of these $|N|$ eigenvalues contains useful information for interpreting the structure of the graph. The number of zero-valued eigenvalues represent the number of connected components in the graph. Optical networks are always connected and therefore only $\lambda_0 = 0$. The first non-zero component, i.e. λ_1 in our case, is often referred to as the algebraic connectivity [25], indicating the general connectivity of the graph and gives a sense of how bottle-necked the graph is. From a clustering analysis point of view this eigenvalue can be associated with the main cluster of data, with subsequent eigenvalues associated with smaller clusters within the graph. Another metric defined from the Laplacian, is the spectral gap. This is defined as the difference between the two largest eigenvalues, i.e. $|\lambda_{|N|}| - |\lambda_{|N|-1}|$. Often dimensionality reduction techniques are used to interpret this distribution better [86]. However, the entire structure of the graph is represented within this distribution and therefore global structures can be interpreted, by manipulating these terms further.

If one defines the matrix $L_D = I - B$, then the entries of B are given by Eq.(3.7).

$$(D^{\frac{1}{2}} A D^{\frac{1}{2}})_{ij} = \frac{A_{ij}}{\sqrt{d_i}\sqrt{d_j}} \tag{3.7}$$

As $B = I - L_D$, one can write its eigenvalues as $1 - \lambda_i$. Taking the trace $t(\lambda(B^N))$ of B^N, which is also equal to the sum of the eigenvalues of that matrix, i.e. $\sum_i (1-\lambda_i)^N$. The (i,j) B^N are 0 unless i and j are adjacent, meaning that B^N is the sum of products resulting from $b_{i_0,i_1} b_{i_1,i_2} \ldots b_{i_{N-1},i_N}$, where $i_0 = i$ and $i_N = j$. Now lets define the V-cycles within the graph G described via the sequence $n_1, n_2, \ldots, n_V$ with n_i and n_{i+1} being adjacent to each other and n_V and n_1 too. Using this one can equate the sum of eigenvalues to that of the number of V cycles normalised by the product of their degrees as described in Eq.(3.8).

$$\sum_i (1-\lambda_i)^V = \sum_V \frac{1}{d_{n_1} d_{n_2} \ldots d_{n_V}} \tag{3.8}$$

This metric is formally defined as the weighted spectrum [86].

$$W(G;V) = \sum_{i \in N} (1-\lambda_i)^V \tag{3.9}$$

The right hand side of Eq.(3.8) can also be traced back to random-walk analysis [86]. A random walk starting at node n_i with degree d_i, will choose a next edge with probability $\frac{1}{d_i}$. This means that the probability of a random walk starting and ending at a node is the sum of all the V-cycles. Furthermore, this can also be compared to the clustering coefficient of a network. The clustering coefficient of a network measures

how well on average a node's neighbourhood is connected. This can be de ned as the average number of triangles over the total number of triangles. The clustering coef cient therefore normalises each triangle according to the total possible number of triangles, whilst the weighted spectrum with $V = 3$, normalises each triangle according to the product of its degrees. As two networks can have the same clustering coef cient, however have quite different structures, this is not the same for the weighted spectrum, which can only have the same elements with networks that are isomorphic.

One can see that the weighted spectrum gives some interpretability to the raw spectra of a graph, by correlating the spectrum to some structural features like number of cycles, random walk and clustering coef cient. Another feature of this spectral manipulation is that it dampens the values of eigenvalues close to one. This is useful since the eigenvalues falling within this region are associated with random links in the network and not necessary useful for the overall structure of the network.

The weighted spectrum can be generalised to give a distribution over h bins, to use equal length distributions for different size graphs. This gives a weighted spectral distribution (WSD) as de ned in Eq.(3.10). Eq.(3.10) is useful for larger graphs, for which the computation of eigenvalues is laborious. In this case, an estimation of the eigenvalues falling into the set of bins H can be made, rather than calculating all eigenvalues.

$$WSD(G) ! f \quad h 2 H : (1 \quad h)^{V} f (\quad = h) g \tag{3.10}$$

The weighted spectrum can also be de ned to give a distance between two networks or a distributions of networks as in Eq.(3.11). This is useful to compare graph structures to see similarities between these. In conclusion, the weighted spectral distribution gives an abstract representation of the graph, which is connected to random-walk analysis, clustering coef cients and number of cycles present in the network. It is especially useful to compare the graphs with each other and measure the difference /similarity in structure with Eq.(3.11).

$$F(G_1; G_2; V) = \sum_{h2H} (1 \quad k)^{V} (f_1(\quad = h) \quad f_2(\quad = h))^2 \tag{3.11}$$

By calculating the degree distributions and diameters of graphs, one can give quick and interpretable graph theoretical representations of structures in graphs. However, learning representations of graphs de ned by the spectrum of the graph give insights into global structures of the graph. De ning the WSD, and weighted spectral distance makes the spectrum of graphs more interpretable and gives insight into differences between graph structures. These tools give insights into what graph structures are

present in optical networks and whether a certain generative graph model gives structures comparable to those in optical networks. The next section introduces the generative graph models investigated.

3.2 Generative Graph Models

To study the structure of topologies, we need to not only be able to de ne and measure these, however also have models behind generating these structures. This long has been a research topic of multiple disciplines. The prime motivation behind these models is to investigate statistical properties of graphs, rather than investigating a speci c single graph's nodes and edges. The importance behind this is that one can generalise ndings and discover relationships, rather than just investigating particular topologies. This is especially important in optical networking, where it is common to use old and often few networks to investigate algorithms or technology performance. The impact of structural and physical properties of networks on their maximum achievable throughput is of particular interest here.

The rst appearance of modelling networks as statistical entities, rather than just deterministic entities appeared with the idea of generating random graphs. Given a set of nodes N and some probability p, each edge is sampled as a coin toss to exist or not with probability p. This is the simplest of models and appeared in the 1950s and is referred to as the Erdös-Rényi random graph model (ER). The rst appearance of this model however can be more accurately attributed to Russian concert-pianist, turned mathematician, Anatol Rapoport in his "Connectivity of Random Nets" 1951 paper with Ray Solomonoff [87]. Using this model they discover that the average component size of a network is highly reliant on the average degree $\langle k \rangle$ of the network. When $\langle k \rangle < 1$ the network is partitioned into many smaller connected components. If $\langle k \rangle > 1$ then a large component in the graph starts to connect. These were the rst initial discoveries in connectivity of networks and phase transition of random networks.

In the late 1950s and 1960s research on random networks continued and many papers were published. Paul Erdös and Alfréd Rényi were the most in uential in this context and continued this work in a pure mathematical context with their 1960 paper investigating the structure of random graphs as connectivity increases [88]. Although the study of random networks was hugely important to forming the bases of many generative graph models that we have today, they lack applicability in many domains. In practice they are applicable to ideas in biology and epidemic analysis, however many networks such as social networks and communication networks were argued not to share the properties of random networks [89].

In tandem to this work, much research was performed in analysing real networks, with an interesting observation over several domains. Over citation networks, the world

wide web network, the internet, metabolic networks it was observed that their degree distributions followed that of a power law as in Eq.(3.12). The signi cance of this is that the power law degree distribution emerges in many different real-world networks, meaning that there might be similar generative mechanisms that form these networks.

$$P(d) \quad d^{\ pl} \tag{3.12}$$

Albert-László Barabasi and Réka Albert proposed in 1999 that there is a common mechanism that lies behind these powerlaw degree distributions, later to become known as the *Barabasi-Albert model* (BA) [84]. They made the connection that many networks have a power law degree distribution. Following this observation they proposed that models are not static entities and their mechanism of formation is sequential, i.e. dynamic in nature. Using this idea they proposed a model that grows dynamically to produce degree distributions that follow a power law and term these scale-free networks. Their model starts with two connected nodes, after which new nodes are added in sequentially and nodes obtain new edges in proportion to how many edges they already have. This is called *preferential attachment* and is the underlying principle behind scale-free networks. This causes the network to have a few very highly connected hubs, however the majority of nodes end up being of low degree. This probability for a new edge $p(i; j)$ can be de ned as in Eq.(3.13), which normalises the degree of node j (d_j) by the degrees in the whole graph.

$$p(i; j) = \frac{d_j}{\sum_{k \in N} d_k} \tag{3.13}$$

The preferential attachment model was novel in its conception, fusing research between empirical network study and that of random networks. It understood that many networks are formed dynamically rather than statically and that their degree distributions often follow a power law and that this property is common. This model in some aspects lends itself well to optical networks: (i) they are communication networks (ii) they are dynamic as they expand over time adapting to new demands (iii) highly connected hubs are advantageous from a wavelength-connectivity point of view. However, there is one important limitation in these models, namely that they do not incorporate distances or geography into the formation of the structure. From a modern-day optical networks point of view, these physical properties are signi cant, since they determine achievable information rates that can be communicated across the network. However, there have been other, less celebrated works that have incorporated this into generative models.

The simplest way to include distances into the generation process is to start with a random graph such as the ER graph generation process and incorporate distances in

some manner into the probability of each edge. This was the inception of Bernard Waxman in 1988 [90], who used this model to test routing algorithms for multipoint connections in packet-switched computer networks. This model is later referred to as the Waxman model within literature. He proposed to simply scatter nodes across a rectangle and then gives an altered probability $p(i,j)$ from the ER model as defined in Eq.(3.14).

$$p(i,j) = \alpha_{wax} \exp\left(\frac{D_l(i,j)}{\beta_{wax} L}\right) \tag{3.14}$$

Here α_{wax} and β_{wax} are parameters in the range $(0,1]$ and $D_l(i,j)$ is the distance between nodes i and j and L_{max} is the maximum distance between any two nodes. Larger values of α_{wax} result in networks with higher connectivities, whilst smaller values of β_{wax} result in networks with a higher density of short edges compared to long ones. Within the original paper results are demonstrated across a wide range of α_{wax} and β_{wax}, this demonstrated a good practice of evaluating algorithmic performance in 1988 when compute power was a lot more constrained.

The Waxman model is a simple addition to the ER model to include distances in graph generation. However, the model has no intuition of preferential attachment or scale-free behaviour, logical since it was before the realisation of these trends. Secondly, connected to the first point, is that it is static and has no notion of growing dynamically as many networks do. Furthermore, the throughput of a lightpath within an optical network approximately scales with $\log(\frac{1}{n_{sp}})$, where n_{sp} is the number of spans in the network. Therefore, it is not immediately clear whether exponential distance scaling is applicable to optical networking.

The idea of taking physical locations of nodes into account and to generate connections according to these locations is something that originally was investigated in the context of geographic variation studies in Biology. More specifically the problem was referred to as the regionalisation problem [91]. The graphs resulting from this investigation were referred to as geographic connectivity graphs. Different schemes of connectivity were established, however the most renown of these, outside of the area of geography, came to be known as the Gabriel connectivity [92]. Gabriel graphs, the graphs resulting from the Gabriel connectivity, can be described as a geometric generative model. This model is deterministic, i.e. for a specific set of node coordinates it produces a single unique graph. If the distance between node i and j is denoted by $D_l(i,j)$, an edge $e(i,j)$ exists, if no node lies within a radius $R = \frac{D_l(i,j)}{2}$, drawn from the half way point between the two nodes, illustrated in Figure 3.3. Therefore, if an edge exists or not, is not a statistical property here, there is no underlying distribution, it purely depends on node locations. This model results in strictly planar graphs, which means they have no overlapping edges and therefore are

Fig. 3.3: Demonstration of gabriel graph generation, where an edge is added if there is no node closer than radius R.

highly grid-structured. This property is something that can be observed in optical networks and generally is desirable [25]. The lack of flexibility makes this model less suitable for large scale network investigation, as connectivity properties of these networks rely only on the geographic distribution of the nodes.

Therefore, there is still a lack of generative graph models that include distance, dynamic generation and specific degree distributions. At the start of research described in this thesis there were several pieces of work that investigated the modelling of optical core networks through generative graph models. Due to the limitations of these investigations, an extension of the BA model for optical core networks was developed and is presented in the following section.

3.3 The Signal-to-Noise Ratio-Aware Barabasi-Albert (SNR-BA) Model

Most papers investigating algorithmic performance used specific (and a relatively small number of) published 'real' network topologies as benchmarks for testing algorithms and performance metrics [77, 79, 58, 83, 82, 80, 81, 78]. These are not sufficient for the evaluation of routing algorithms or network performance, as it is

important to con rm that the performance is due to an algorithmic improvement and not due to some topological artefact.

Previously, the evaluation of wavelength requirements has been the focus of such studies. Initially random topologies were investigated. Fenger looked at the impact of topology on wavelength requirements, by generating millions of topologies randomly and evaluating them in terms of node degree variance and showed that the number of spanning trees in a topology is a good indicator as to the wavelength requirement of a network [93]. The problem here was that only randomly generated networks were considered and it is well known that these networks give structures not necessarily observed in real networks [84]. Therefore, only investigating the parameters on these structures is not conclusive. This work was then later picked up by Châtelain, who looked at randomly generating networks, constraining these by geographical distances [26]. These networks were then used to understand which graph theoretical properties correlate to wavelength requirements of these graphs. He concluded that algebraic connectivity (rst non-zero eigenvalue of the Laplacian of a graph), correlates the most to the wavelength requirements of networks. Again the generative graph model was not investigated as to what structures were produced and only these structures were investigated. Yuan picked up this thread and identi ed that regular structured topologies are not representative of optical networks and therefore looked at wavelength requirements within graphs generated by both BA and Watts-Stogratz (WS) models, exhibiting scale-free and small-world behaviour respectively [89]. A conclusion here is that wavelength requirements are realised to be correlated to the average path lengths of graph, which is a result directly related to work done by Baroni and Bayvel in 1997, which showed a direct relationship between network connectivity and wavelength requirements [18].

As optical ampli ers led to the expansion of network throughput via huge WDM systems, physical properties of these systems became the limiting factor for throughput. Before the work in this PhD research, there was little work, that has investigated the impact of both structure and physical properties of a topology on the maximum achievable throughput of networks. It is important to be able to generate structures and physical properties close to those of real optical networks, to understand the impact of each. This is key to the understanding of how networks should be designed and what is limiting them.

Most of this previous work using generative models did not include physical properties of optical networks; these are known as non-geometric generative models [94, 95, 96, 97]. Although [98] analysed the maximum achievable nonlinear throughput for a range of networks using their physical properties, the process of topology generation followed that of non-geometric generative models, not adequately re ecting spatial information.

Recently, [7] included physical properties in the generation of topologies using a genetic algorithm, to quantify maximum achievable nonlinear throughput. With the main focus on the comparison of heuristic versus ILP solutions for resource allocation, this work, however, did not include the necessary analysis of the generated graph structures.

There was a number of developments made to include physical properties in the graph generation process in geometric generative graph modelling - highlighted in section 3.2. Pavan et al. [99] modelled optical core networks via a geometric generative graph model, which scales the probabilities of edges via an exponential decay (known as the Waxman model). However, the exponential decay of edge probabilities with distance is not representative of the distance-dependent penalty observed in optical bre networks. Other geometric generative models, in turn, failed to model well-connected local hubs [25], typical of optical core networks connecting major centres. In [100] graphs were generated using both Waxman and Gabriel graph models, however examined neither the structure of these graphs nor the physical properties when investigating their performance.

With this previous work in mind, there still was lacking a generative graph model that takes these features into account:

(i) distance dependent penalties that scale according to optical network transmission

(ii) close representation of structures compared to real optical core networks

Physical properties of the optical bre and ampli ers used in transmission, as well as link (edge) lengths, de ne the transmission performance of optical networks. Thus, certain edges may be structurally bene cial, i.e. enabling more lightpaths to be setup, however these may be heavily physically impaired. To include the physical properties of links, and their effect on data transmission and network throughput, requires the addition of a metric which describes their impact within the network generation. The metric chosen in this work is the SNR of a given transmission link (or edge) and it is included within the probability function when choosing edges in the graph generation process.

As described earlier, since geometric generative models can capture the grid-like behaviour of real optical core networks, yet fail at modelling local hubs, in the research work described in this thesis, it was proposed to extend the probability weights of the conventional BA model, as de ned in Eq.(3.13). The proposal was to include the SNR between any two given nodes, as a second weight in the conventional BA model. The SNR term attempts to make realistic link decisions via weighing shorter links more heavily and the BA term attempts to replicate local hubs in the network, demonstrated to be important in [25]. A weighting parameter η is used to determine how heavily

to weight the physical properties within the graph generation process. The probability weights of the SNR-BA are then given by

$$P_{\text{SNR-BA}}(i;j) = \left(\frac{\text{SNR}(i;j)}{\sum_{k \in N} \text{SNR}(i;k)} \right)^{\alpha} \cdot \frac{d_j}{\sum_{k \in N} d_k}; \tag{3.15}$$

where $\text{SNR}(i;j)$ is the SNR on the direct link between nodes i and j. The SNR includes the effects of distortion arising from the optical Kerr effect. The latter can be approximated as noise, referred to as nonlinear interference noise, and amplified spontaneous emission noise from optical amplifiers. Numerous models of calculating the SNR of an optical lightpath have been proposed in the literature. Following one of the most widespread modelling approaches, namely a first-order perturbative description of the nonlinear interference noise, the SNR at optimum launch power is given by [101].

$$\text{SNR} = \sqrt[3]{\frac{1}{\frac{27}{4} P_{\text{ASE}}^2 \eta^3 n_{\text{sp}}^3}}; \tag{3.16}$$

where P_{ASE} is the injected amplified spontaneous emission noise per amplifier, η the nonlinear interference coefficient and n_{sp} is the number of fibre spans. Eq.(3.16) assumes an incoherent addition of nonlinear interference across multiple spans, a common assumption in the physical layer modelling of optical networks which imposes negligible inaccuracies for C-band transmission and beyond, cf. 1) in [101, Fig. 10-11]. Recalling that the number of spans is given by $n_{\text{sp}} = \frac{D_l(i;j)}{L}$, where L is the span length and $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, Eq.(3.16) can be written as

$$\text{SNR}(i;j) = \text{SNR}_1 \left(\frac{L}{D_l(i;j)} \right)^{\frac{3}{2}}; \tag{3.17}$$

where SNR_1 is the SNR after a single span. Substituting Eq.(3.17) in Eq.(3.15) yields the proposed probability weights as

$$P_{\text{SNR-BA}}(i;j) = \left(\frac{\frac{L}{D_l(i;j)}}{\sum_{k \in N} \frac{L}{D_l(i;k)}} \cdot \frac{d_j}{\sum_{k \in N} d_k} \right)^{m} \cdot \left(\frac{1}{\sum_{k \in N} \frac{D_l(i;j)}{D_l(i;k)}} \cdot \frac{d_j}{\sum_{k \in N} d_k} \right)^{l}; \tag{3.18}$$

where the approximation is introduced by dropping the rounding operation. The derivation of Eq.(3.18), assumes that the amplified spans have identical lengths throughout the network. While this is not always satisfied in practice, Eq.(3.18) still describes the average SNR scaling with respect to distance.

By using the BA model as a base, a dynamic building process of the graph is included, however derive a distance scaling that is based on the average SNR scaling in optical networks. How well this model matches the strutctures of real optical core networks is investigated next.

3.4 Real Optical Core Networks

The goal of validating a generative graph model that is close to that of real optical core networks is two-fold: (i) to allow for generating set of networks for testing algorithmic solutions to network design and resource allocation for the future and (ii) to investigate the structures and physical properties present in real-optical networks.

Therefore, a data set of 25 core optical network topologies from the survivable network design library (SNDlib) [102] and CONUS topologies [8], are used in the comparison with the graphs generated via the models introduced in section 3.2, and are summarised in Table 3.1. NSFNET is excluded here, as it was not part of the SNDlib. Here the connectivity refers to the connectivity defined in Eq.(2.19). The constraints on the choice of topologies for the set was that no network graph could be cut in two by removal of a single edge, to ensure resilience. Although most networks used are legacy networks, designed in an era where optical networks were opaque, the design goals within opaque and transparent optical networks are very similar. Opaque networks are networks that regenerate the optical signal at each intermediate node in the network. Therefore, opaque networks aim at minimising edge lengths and the diameter of the network, as to minimise the number of regenerations required. Networks today, are transparent, meaning that they do not regenerate the lightpath at each node. Transparent networks aim at minimising the path lengths within the

network, to minimise physical layer impairments [103].

This dataset allowed for accurate distance modelling using exact geographical node locations for each network. To model distances in these graphs, their geographical coordinates were used in conjunction with the Haversine formula [104]. The Haversine formula takes into account the curvature of the earth and calculates distances over a sphere rather than a plain. Realistic bre distances were also accounted for by using Eq.(3.19), where D_{hav} represents the Haversine distance. The bre distances are estimated according to the European Telecommunications Standards Institute (ETSI) guidelines [105].

$$D_{fibre} = \begin{cases} 1.5\ D_{hav} & \text{if } D_{hav} < 1000\text{km} \\ 1500\text{km} & \text{if } 1000\text{km} \leq D_{hav} \leq 1200\text{km} \\ 1.25\ D_{hav} & \text{if } D_{hav} > 1200\text{km} \end{cases} \tag{3.19}$$

Network	jN j	jE j	Connectivity (!)
Abilene	12	15	0.227
Atlanta	15	22	0.209
Cost266	37	57	0.085
France	25	45	0.150
Geant	22	36	0.155
Germany50	50	88	0.071
Giul39	39	86	0.116
india35	35	80	0.134
janos-us-ca	26	42	0.129
janos-us	39	61	0.082
nobel-eu	28	41	0.108
nobel-germany	17	26	0.191
nobel-us	14	21	0.230
norway	27	51	0.145
pioro40	40	89	0.114
polska	12	18	0.272
sun	27	51	0.145
ta1	24	51	0.1845
ta2	65	108	0.051
zib54	54	80	0.055
CORONET-CONUS	75	99	0.035
CORONET-GLOBAL	100	136	0.027
CORONET-CONUS-60-ONDP	60	77	0.043
CORONET-CONUS-30-ONDP	30	36	0.082
CORONET-CONUS-60	60	79	0.044

Table 3.1: Table of real networks used in structural comparison of generative graph models.

These networks are used as a base to compare against the generative graph models

introduced in section 3.2. They are then compared for similarity by comparing the degree, diameter and spectral properties as described in section 3.1. The following section describes the methodology and presents this comparison between generative graph models and real optical core networks.

3.5 Comparing Generative Models and Optical Core Networks

A set of 200 graphs were generated for each of the 25 real optical core networks, introduced in section 3.4, by taking the node-coordinates of the real networks, and generating the set of graphs following the rules of each of the generative models of section 3.2, except for the Gabriel graphs for which only a single graph was generated, as this model is deterministic in nature and not statistical. In future simulation studies, node locations can be generated by uniformly scattering nodes on a grid of a given distance scale, as done in [90] and in Chapters 4 and 5 of this thesis. A radius of length r can be de ned to ensure no two nodes are closer than this distance, such that nodes are spaced evenly across the grid. This is because optical core networks span large (100/1000s km) areas and, therefore, minimum distances should be enforced, so that unrealistic transmission cases do not arise.

In total 5000 graphs were generated by each generative model. In the case of the ER graphs, the graphs were rst created and then the nodes assigned positions, as the node positions had no effect on the creation of the graph. In addition, since the BA graphs did not take into account distances, the coordinates did not matter, however the order in which one added nodes to the graph did, as nodes added early on in the sequence tend to act as highly connected hubs. This comes down to the rich-gets-richer principle, where the longer a node is in the generation process, the higher probability and chances it will have to attract more edges. Additionally, with each additional edge, a node will attract more edges. Since the node sequence affects the graph generation, the same method for selecting the sequence of nodes for the sequential generative models (BA, Waxman, SNR-BA) was followed. The initial node selected in the sequence was the most central of all the nodes in the graph. This was determined rstly by nding the centroid of the graph via Eq.(3.20), where x and y refer to the latitude and longitude, respectively. The node with the smallest Euclidean distance to the centroid at the beginning of the sequence was added rst. After this the next nodes were chosen sequentially by comparing every other node's average distance to all other nodes already present in the node sequence; each time adding the node with the smallest average distance to all other nodes in the sequence.

$$C_{x;y} = \left(\frac{x_1 + x_2 + \cdots + x_n}{n}; \frac{y_1 + y_2 + \cdots + y_n}{n} \right) \tag{3.20}$$

The β parameter in Eq.(3.14) and Eq.(3.18) (the distance weighting factor) was

Model	WSD	Degree		Diameter		Spectrum	
	F	D_{KS}	p_{KS}	D_{KS}	p_{KS}	D_{KS}	p_{KS}
BA	0.00045	0.190	0	0.468	0	0.0580	0.003
ER	0.00087	0.222	0	0.397	0	0.0541	0.008
SNR-BA	0.00019	0.035	0.177	0.119	0.827	0.0120	0.997
Wax	0.00020	0.067	0	0.343	0.004	0.0352	0.194
GG	0.00067	0.141	0	0.252	0.327	0.0553	0.104

Table 3.2: Weighted spectral distances (F), Kolmogorov-Smirnov two sample test statistic (D_{KS}) and p-value (p_{KS}), calculated for the degree, diameter and spectra of the graphs generated by ER, BA, Waxman, GG and SNR-BA models.

selected via a simple sweep of the parameters to obtain values that performed well as a starting point, after which a non-gradient based optimisation was applied using the Nelson-Mead method [106]. The cost function used to evaluate the parameters was the weighted spectral distance, $F(\cdot)$, as defined in Eq.(3.11). It was found that the weighted spectral distance between the real graphs and those generated by SNR-BA was smallest when using $\alpha = 5$. For Waxman graphs the smallest weighted spectral distance occurred when using $\alpha_{wax} = 8.5$. These values were used to generate the final set of graphs used for the analysis in this paper.

For the ER graphs p was chosen by evaluating $p = |E|/\binom{|N|}{2}$, where E and N are the sets of edges and nodes in a graph. If a graph was not bi-connected, then p was incrementally increased by 0.01 and another graph generated. This is repeated until a feasible graph was found; on average 16.7% of the graphs generated were feasible graphs. For the BA model m was chosen to be $\left\lfloor \frac{|E|}{|N|} \right\rfloor$, after which edges are added to nodes until the desired number of edges was reached.

For each of these datasets, their degree, diameter, spectra and WSDs were calculated as described in section 3.1. These distributions were then tested against the distributions of real topologies of section 3.4 using the Kolmogorov-Smirnoff (KS) two sample test. This yields two key metrics: a statistical distance metric between two distributions, D_{KS}, and a probability, p_{KS}. D_{KS} gives the largest variation between any two distributions. Smaller D_{KS} values indicate a small variation between the two distributions. The probability p_{KS} indicates whether the two distributions originate from the same population, with values close to one showing this, and is formally defined as the probability of a sample of size n having a D_{KS} of less or equal than the observed sample [107]. The weighted spectral distance, as defined in Eq.(3.11) indicates how different the WSD of two distributions are, where smaller values show greater similarity. The results of the KS two sample tests and the weighted spectral distances are summarised in Table 3.2.

3.5.1 Generated Degree Distributions

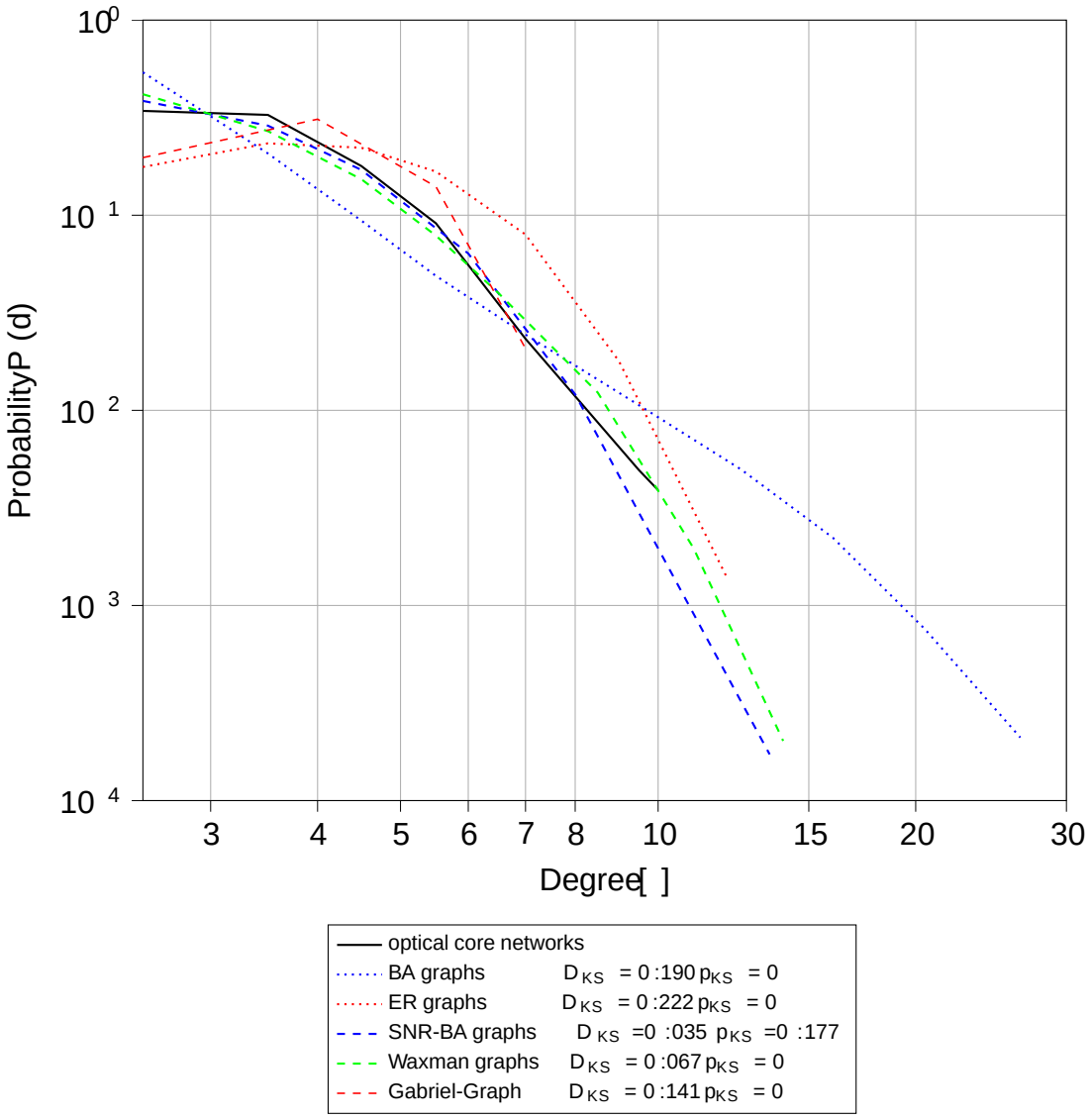


Fig. 3.4: Degree distributions for real optical core networks and graphs obtained from the ER, BA, Waxman, Gabriel-Graph and SNR-BA models.

The degree distributions were calculated for all the generated graphs and are plotted in Figure 3.4. The black curve shows the distribution of the real optical core networks and its important to notice that this distribution shows that the optical networks do not necessarily display scale-free behaviour in terms of their degree distribution. This would otherwise be close to that of the BA graphs, which exhibit power law behaviour until the tail-end of the distribution, where the distribution is skewed due to the larger proportion of smaller graphs in the real optical core networks data set.

The ER graphs yield degree distributions closer to Poisson distributions. This is because the ER graph generation aimed to create networks with the same number of edges as the sparse real optical core networks, and the resulting degree distribution has

a sharper drop-off. Although the ER degree distribution look similar to that of real optical core networks, the beginning and end of the tail of the distribution differ quite signi cantly, giving a much narrower degree distribution.

Finally comparing the SNR-BA, Waxman and GG models, it is clear these distributions t more closely to the distributions of the real optical core networks, although still creating graphs with higher degrees than real optical core networks. The SNR-BA graphs match the start and end of the tail better than those of the Waxman/GG, and this is con rmed in Table 3.2, where the KS two sample test for the degree distributions of the SNR-BA graphs produces the smallest absolute distance to the real optical core networks and the largest likelihood p_{KS}, indicating a larger probability that the two distributions originate from the same population than that of the other generative models.

The SNR-BA model, thus, appears to generate graphs with degree properties closest to the real optical network topologies. This gives us an indication that utilising a process that still incorporates preferential attachment and a sequential process still is bene cial to creating the desired degree distributions. Its important to look deeper at the structure of these networks and understand what are the inherent structures that give these degree distributions and are the structures similar to those present in real optical core networks. To do that, the distribution of diameters given by these models are analysed in the next section.

3.5.2 Generated Diameter Distributions

The diameter, as described in section 3.1.2, was calculated for each of the 20,050 graphs and the distribution plotted in Figure 3.5. The real optical core networks exhibit long diameters with a large range of up to 16. This can be explained by the linear scaling of both the nodes and edges in the network, therefore, networks becoming more sparse as they grow in size. In addition, this indicates that the real optical core networks are grid-like in nature [25], leading to these large diameters as node scales increase.

The BA graphs can be seen to have smaller diameters, peaking around a diameter of 4. This is mainly because BA graphs create multiple highly connected hubs that tend to make connections across the graph - regardless of distance. ER graphs also have shorter diameters peaking around a diameter of 5, however as the edges within ER graphs are modelled as independent Bernoulli distributions, they do not create highly connected hubs within the graph, resulting in slightly longer diameters.

In the Waxman graphs, it is easy to see that when introducing distances into the probabilities of edge creation, the diameters of the graphs start to increase. This is best demonstrated by the SNR-BA/Gabriel graphs that have very large diameters, as they prefer shorter edge connections and, therefore, create grid-structures with inherently

longer diameters. Again it is easy to see in the Table 3.2 that the SNR-BA graphs have the smallest D_{KS} value, as well as the largest p_{KS} for their diameters compared to the real graphs. This again indicates that using a mixture of preferential attachment and distance scaling can give us comparable shortest path structures to those in optical core networks.

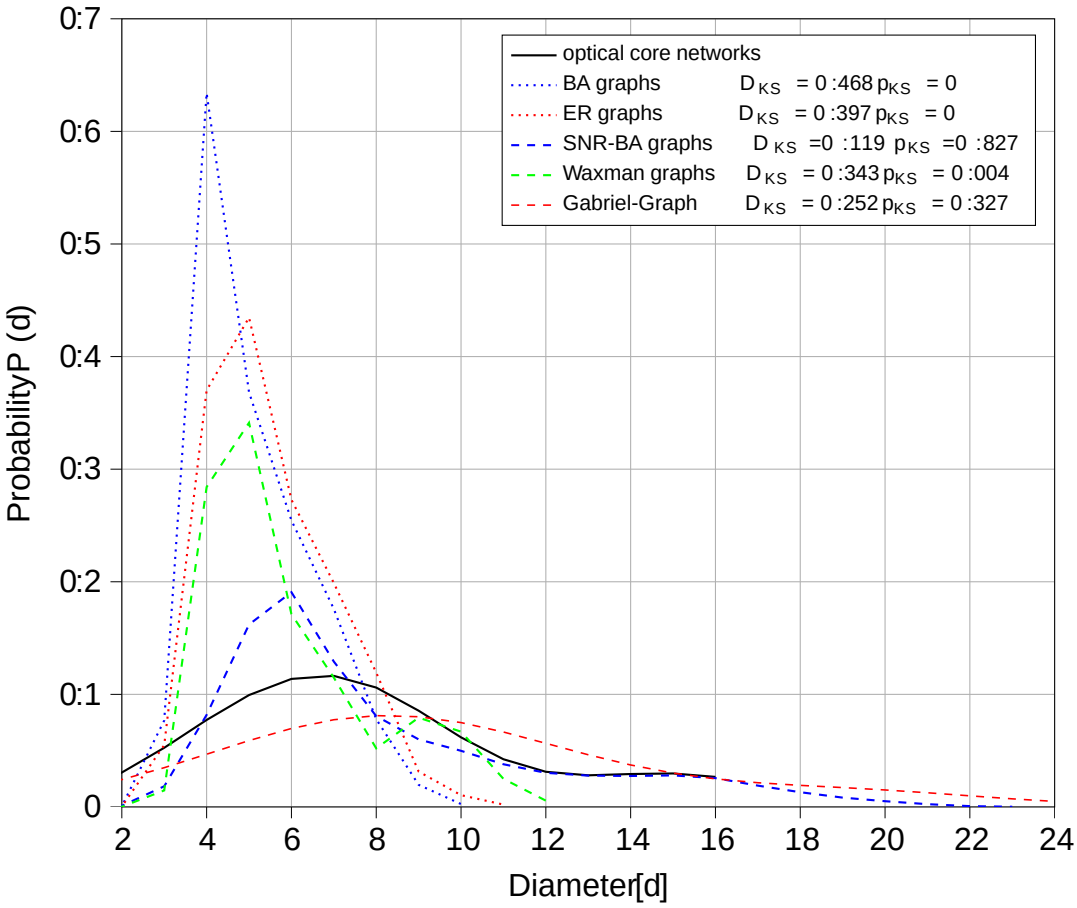


Fig. 3.5: Probability distributions of diameters for real optical core networks and graphs obtained from ER, BA, SNR-BA, Waxman and Gabriel Graph generative models.

3.5.3 Generated Spectral Properties

The spectra of the eigenvalues of the normalised Laplacian were compared to give greater insight into the structure of the generated graphs, following section 3.1.3. As discussed previously, the key regions are the peaks around the smallest and largest eigenvalues. The WSD of all 20,050 networks was calculated and plotted in Figure 3.6.

Examining the behaviour of the peaks on the left hand side, it can be seen that the SNR-BA graphs best re ect the smaller algebraic connectivity values of real network topologies, followed by the Waxman graphs. This can be interpreted as these graphs share more common clustering of nodes and structures. The BA and ER graphs tend to

have the peaks of their spectrum closer to the centre, showing (a) algebraic connectivities are larger and (b) more of their eigenvalues are clustered around the centre of the spectrum, around the value of 1. On the right hand side, for eignevalues larger than 1, the BA graphs exhibit very similar bi-partite structure compared to SNR-BA graphs, shown by the peaks at similar positions. It is the Waxman graphs which appear to mimic the strong bi-partite nature of real optical core networks the best, indicated by the overlapping peaks, followed by the SNR-BA and then ER/GG graphs. However, the SNR-BA graphs have the smallest weighted spectral distance (F), as well as lowest D_{KS} and largest p_{KS} values of all generative models, and are thus, the most similar in structure to the real optical core networks. This comparison of the distributions of the generative models and the relative differences to the real optical core networks, highlights the importance of incorporating distances when modelling optical core networks. The proposed SNR-BA model combines the key structural features of optical network topologies together with a description which re ects their physical behaviour.

In conclusion, Table 3.2 shows that the SNR-BA networks match the set of real

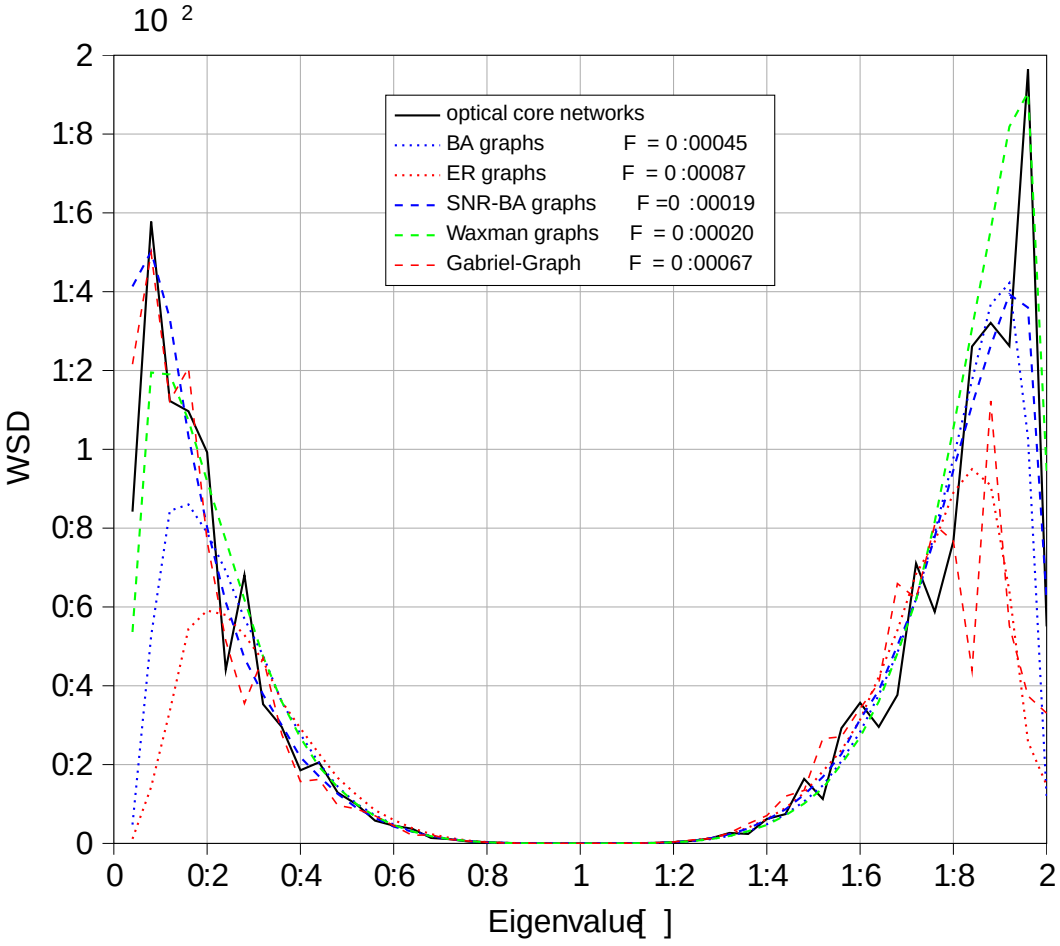


Fig. 3.6: Weighted spectral distribution using $\gamma = 4$ for real optical core networks and graphs obtained from all generative graph models.

networks the best according to both Kolmogorov-Smirnov two sample test statistics for degree, diameter and spectrum and have the smallest weighted spectral distance to the real networks. This analysis can now be applied to investigate the relationship between the network structure, physical properties and performance, in terms of key optical network performance metrics, namely wavelength requirements and throughput. This is carried out in the next section using two sample optical network topologies, CONUS and NSFNET, focusing on the comparison to geometric (SNR-BA) and non-geometric (ER, BA) graphs. CONUS was selected as a representative north-American topology from the set of 25 networks used and NSFNET has been widely used in literature as a sample topology for a variety of studies [18, 75, 7]. Waxman and GG graphs were excluded to reduce computational time, as SNR-BA graphs were shown to give more structural similarities to real optical core networks when including distances in the edge creation process. Furthermore, GG graphs are deterministic and, therefore, cannot produce any statistically signi cant results in this case study [92].

3.6 Comparing Structural and Physical Properties of Optical Core Networks

Both the wavelength requirements and maximum achievable throughput of networks are analysed to understand the affect of structural and physical properties of networks on the performance of networks. Generative graph models can be used to generate networks with various structures and physical properties, as seen in gures 3.4, 3.5 and 3.6. By generating these different topology structures and physical properties, we can investigate their impact on optical network performance and understand better how to incorporate this into future optimisation schemes.

The 30-node CONUS topology and the NSFNET [108] were used for the node locations of this investigation. The NSFNET was originally designed in the 1980s and focused on connecting supercomputer centres, not major city sites, and it is not clear whether it is still an applicable topology for analysis. However, it contrasts more modern networks and has been widely used in previous research studies and therefore included here. Starting with the CONUS and NSFNET node positions, 200 graphs were generated by ER, BA and SNR-BA, respectively, with a total 600 graphs per topology. The constraint imposed on the generated topologies was that no graph could be cut in two by the removal of a single edge (bi-connectivity). As the CONUS topology is very sparse (connectivity, de ned in Eq.(2.19) of 0.082), the ER and BA graphs struggled to satisfy the resilience constraint, creating graphs with 25% and 19% more edges than the original network. The graphs based on the NSFNET node locations have exactly 21 edges, like the original NSFNET.

3.6.1 Wavelength Requirements

An ILP based on [109] was used to calculate the wavelength requirement (w_r) of the graphs. For a network, a set K_z of k-shortest paths was found using Yen's algorithm [36], which iteratively finds alternate routes between source and destination nodes of varying lengths. The value of $|K_z|$ was set to 100, although in most cases a significantly smaller number was achieved and used; the resultant paths were filtered so that only paths of the same lengths (number of hops) as the shortest path were used. This was done to reduce number of paths used in the routing and therefore reduce the complexity of the ILP to solve. In addition, number of hops was used, as it is more likely there exist multiple paths of same number of hops, compared to multiple paths having exactly the same physical distance. A set Z of node pairs (equipped with a set of maximum W wavelengths) needs to be connected via a RWA via a set K_z of paths. The decision variable $_{wkz}$ - with $w \in W$; $k \in K_z$; $z \in Z$ - is able to fully define the RWA of a network following the definition in Eq.(3.21).

$$_{wkz} = \begin{cases} 1 & \text{if } (k, w) \text{ are the RWA assignment} \\ & \text{for node pair } z \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

To define whether a wavelength is needed for routing within the network, the variable w_o is defined as in Eq.(3.22).

$$w_o = \begin{cases} 1 & \text{if wavelength } w \text{ is used in any} \\ & \text{assignment in } _{wkz} \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

Constraining $w_o \geq _{wkz} \; \forall w \in W \; \forall k \in K_z \; \forall z \in Z$.

Using w_o, an objective function is defined to minimise the sum of w_o over all $w \in W$, as defined in Eq.(3.23).

$$w_r = \min \sum_{w \in W} w_o \tag{3.23}$$

The ILP solution set needs to be constrained, so as to only give solutions that are feasible for optical networking. Firstly, by only assigning a single path and wavelength per node pair $z \in Z$ as defined in Eq.(3.24).

$$\sum_{w \in W} \sum_{k \in K_z} _{wkz} = 1 \quad \forall z \in Z \tag{3.24}$$

Secondly, no two node-pair path assignments can share a wavelength on any given edge

j. Therefore, the variable $I_e(j \in k)$ is defined to be 1 when edge j is in path k and 0 otherwise. Using this the wavelength uniqueness can be constrained as in Eq.(3.25).

$$\sum_{z \in Z} \sum_{k \in K_z} w_{kz} I_e(j \in k) \leq 1 \quad \forall j \in E \quad \forall w \in W \tag{3.25}$$

The ILP yielded the minimum wavelength requirements, that is the minimum number of wavelengths needed to route the $N(N-1)/2$ demands between all node pairs, determined by the objective defined in Eq.(3.23). The wavelength requirements for the 1200 ER, BA and SNR-BA networks based on the NSFNET and CONUS networks are calculated and plotted alongside the wavelength requirement of NSFNET and CONUS in Figure 3.7. For the CONUS network 122 are required, for the NSFNET this number is 13, same as determined in [18]. The box-plot shows the distribution of the data together with the median, interquartile range and the minimum and maximum values.

Figure 3.7 shows that the CONUS-based ER and BA graphs have 52% and 51% lower wavelength requirements than the SNR-BA graphs. Similarly, NSFNET-based ER and BA graphs have 31% and 23% lower wavelength requirements, than the SNR-BA graphs. The ER and BA graphs appear to have a structural advantage, in terms of wavelength requirements, over the SNR-BA graphs, because of their smaller diameters and edge connections spanning larger parts of the graph.

Wavelength requirements used to be a key performance metric in networks, due to wavelengths being a scarce resource. From this point of view one can see that, in general, the ER and BA graphs for both the CONUS and NSFNET node locations perform well, whereas the SNR-BA graphs have on average about 130% and 40% higher wavelength requirements than the ER and BA networks for CONUS and NSFNET respectively. The CONUS topology, published more recently and now considered more representative of real networks [8], has a wavelength requirement that agrees more with that of the SNR-BA graphs - only 16% less than the SNR-BA networks on average, compared to 44% and 55% on average for the BA and ER graphs based on CONUS. NSFNET has a lower wavelength requirement ($w_q = 13$) and agrees with the average of the ER networks and only 13% less than the average of the BA networks. Here SNR-BA networks have on average 53% higher wavelength requirements. From this, it is is clear that both ER and BA networks generally give smaller wavelength requirements. It is important to note that some ER networks generated more edges than in the original network, especially for the CONUS edge numbers, since it is a very sparse network. This means that with more edges (higher connectivity) a lower wavelength requirement follows, as shown in Baroni et al. [18]. Indeed, wavelength requirements, are generally only explained by the structural part of the network, neglecting the physical properties in the network. However, the

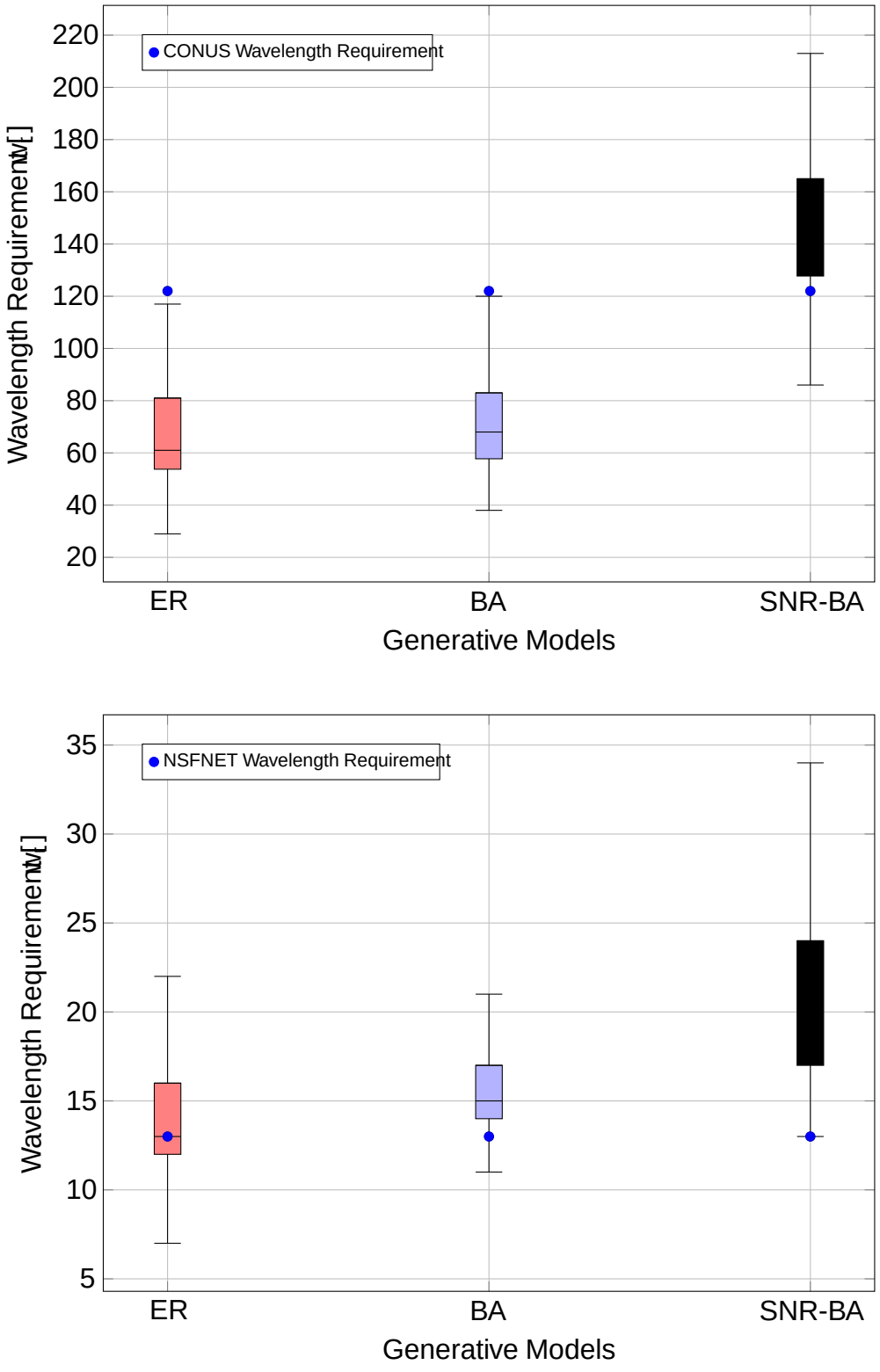


Fig. 3.7: Wavelength requirements shown for the CONUS and NSFNET, as well as box-plots illustrating the distribution of wavelength requirements for the graphs generated by ER, BA and SNR-BA models.

overarching goal of network design is to maximise throughput for the entire network, which requires physical transmission impairments to be taken into account.

3.6.2 Maximum Achievable Throughput

Linear and nonlinear fibre transmission impairments need to be taken into account to calculate the throughput of optical networks. Unlike the SNR derived in section 3.3, here we do not use optimal launch powers, but apply a uniform launch power of 0 dBm (see for example [29]) for all wavelength channels. To calculate the SNR, a closed form GN model was used [29], where the SNR of a lightpath i is given by Eq.(3.26) as described in section 2.1.2. Here P_i is the launch power of lightpath i, η_{NLI} is the nonlinear coefficient that can be calculated following equation 5 in [29] and P_{ASE} is the power of the amplified spontaneous emission.

$$SNR_i \quad \frac{P_i}{P_{ASE} + P_i^3} \tag{3.26}$$

For this set of simulations, all links were assumed to be multiples of 80 km standard single mode fibre spans with $= 0{:}2 \frac{dB}{km}$, $D_s = 18 \frac{ps}{mm\ km}$ and $= 1{:}2 \frac{1}{W\ km}$. In between spans, erbium-doped fibre amplifiers, each with a noise figure of 4 dB was used (as in [98]), although any practical value larger than the fundamental limit of 3 dB can be used for this analysis. As before, Nyquist-spaced ILP channels of 16 GBd were used. Although filtering effects at the ROADM genereally have roll-off on either side of the passband, this is neglected in this work. They were interfaced with colourless, directionless and contentionless, reconfigurable optical add-drop multiplexers (CDC-ROADM) over a constrained C-band (1530-1570 nm) optical bandwidth. The losses, filtering effects and amplification needs of the ROADMs were not considered in this work.

An ILP based on [75] was used to maximise the nonlinear throughput. This maximises the throughput given a uniform bandwidth across all node-pairs. To solve the ILP, the k-shortest paths were calculated between all node-pairs, after which the SNR for these paths was calculated assuming the worst case SNR: that of a centre channel in a fully loaded link, with 0 dBm launch power per channel, following Eq.(3.26). Using this SNR, their achievable capacity can be calculated via Shannon's theorem, as defined in Eq.(3.30) [6]. Although the Shannon capacity is the upper bound and there are many forward error correction (FEC) codes that represent what is achievable, it is used here to illustrate the difference in achievable throughputs between networks [110, 111]. This set of capacities is referred to as $C_{z,k}$ where $z \in Z$ is a node pair (i, j) and $k \in K$ is a path. $\overline{T_z^B}$ is the normalised traffic matrix, in our case simply kept at uniform across the node pairs and w is a continuous integer variable over which one tries to find an RWA using w_{kz} such that every node-pair is able to route proportionally the same amount of bandwidth determined by $\overline{T_z^B}$. Eq.(3.27) ensures that every node pair routes at least the product of the traffic matrix and the throughput multiplier c and Eq.(3.28) describes the objective of maximising c. All the previous

constraints need to be followed, to ensure a feasible RWA set. Using this ILP the RWA throughput can be maximised given zero-blocking and uniform bandwidth demand.

$$\sum_{w} \sum_{k} w_{kz} C_{z;k} \quad c\overline{T_z^B} \quad 0 \quad 8 \quad z \; 2 \; Z \tag{3.27}$$

$$c_{max} = \max(c) \tag{3.28}$$

The resultant throughput for the RWA was calculated from the accumulated SNR of each lightpath assignment. A lightpath $= f p_i ; w_i g \; 2 \; R$ has a path p_i and a wavelength w_i associated with it and is part of the set of lightpaths for a routing R. To calculate the capacity for this lightpath, one first needs to take into account the edges along which the lightpath travels and their respective SNR values. Using Eq.(3.26) and the state of the network i.e. the knowledge of which wavelengths are present on which links, one can calculate the $SNR_{(i;e)}$ value on each of the links $e \; 2 \; p_i$. The total SNR of that path is then calculated by taking the inverse sum of the NSR (inverse of SNR) values of each link traversed by the path p_i, shown in Eq.(3.29).

$$SNR_i = \sum_{e2 p_i} \frac{1}{SNR_{(i;e)}} \quad ^{1} \tag{3.29}$$

This SNR can then be used to calculate the maximum achievable data rate over this lightpath using Eq.(3.30) [6]. It can be seen that the capacity of a lightpath mainly depends on the SNR of that path, which, in turn, depends on the length and congestion along the path. B_{CH} represents the channel bandwidth used, which is kept constant at 16 GHz for all channels.

$$C_i = 2 B_{CH} \log_2(1 + SNR_i) \tag{3.30}$$

$$T = \sum_{i2R} C_i \tag{3.31}$$

The throughputs for all the lightpaths, allocated to satisfy the demand, were calculated and summed, as in Eq.(3.31) to give the total achievable throughput of the RWA for a particular network.

Although this work was published in [112], there was a small error where the wrong channel bandwidth was used to calculate the throughput. Within this PhD thesis this error was corrected and the correct values were calculated and plotted throughout. The error did not make a change in the conclusions of the work.

The maximum uniform throughput values T) of the ER, BA and SNR-BA graphs based on the CONUS and NSFNET topologies were calculated and are shown in Figure 3.8. It can be seen that it is now the SNR-BA graphs that, on average, perform 59% better than the ER graphs and 54% better than the BA graphs based on the

CONUS topology, despite the greater number of edges in the ER and BA graphs. For the NSFNET-based graphs, the SNR-BA graphs, on average, outperformed the BA and ER graphs by 49% and 30%, respectively. Therefore, it is clear that the ER and BA graphs, on average, perform worse than the SNR-BA graphs for both example networks.

This drop in performance between the ER and BA graphs compared to the SNR-BA graphs is the result of longer path lengths in the former. Figure 3.9a shows the distribution of average path lengths for all the generative models based on the CONUS network and 3.9b for the NSFNET graphs. The paths in the CONUS-based ER and BA graphs are on average 215% and 187%, respectively, longer than those taken over the SNR-BA graphs. For the NSFNET-based graphs, although shorter, the signals travel 95% and 98% further over the ER and BA graphs compared to the SNR-BA graphs. This difference in distances, and the associated transmission penalties, dominate the achievable throughput, and at these distances the structural advantages of the ER and BA graphs do not translate into larger throughputs.

It should be noted that this difference in edge (span) lengths between the SNR-BA graphs and the ER/BA graphs, results in the difference in the total deployed bre lengths. The NSFNET-based SNR-BA graphs use 53% and 47% less in total bre compared to their ER and BA counterparts, respectively. For the CONUS-based SNR-BA graphs, this difference is even larger, saving 72% and 68% of total bre compared to the ER and BA graphs respectively. The SNR-BA graphs are on average able to achieve higher throughput, which directly translates into lower blocking probability, whilst deploying less bre.

In calculating throughput, uniform bandwidth demand was assumed, while a uniform set of connections is assumed when calculating the wavelength requirements. This difference is signi cant as longer lightpaths with lower SNR values, will inherently need to route more connections to satisfy a given bandwidth demand. This difference in the number of connections required to satisfy a given bandwidth demand is driven by the physical properties of the network. This is one of the factors that reduces the structural advantage of the ER and BA graphs, observed when evaluating wavelength requirements.

Due to this, the number of lightpaths established in the RWA is ( ) is not signi cantly different between the SNR-BA graphs and those of ER and BA. Namely, for the NSFNET based graphs, the SNR-BA graphs established on average only 8% fewer lightpaths compared to the ER graphs and 7% more lightpaths compared to the BA graphs. For the CONUS based graphs, the SNR-BA graphs established only 11% and 10% fewer lightpaths, on average, than the ER and BA graphs. This is signi cant, when looking at Figure 3.10, where the distributions of throughput per lightpath ( ) is shown for each generative model. Here the same trend as before can be seen, where

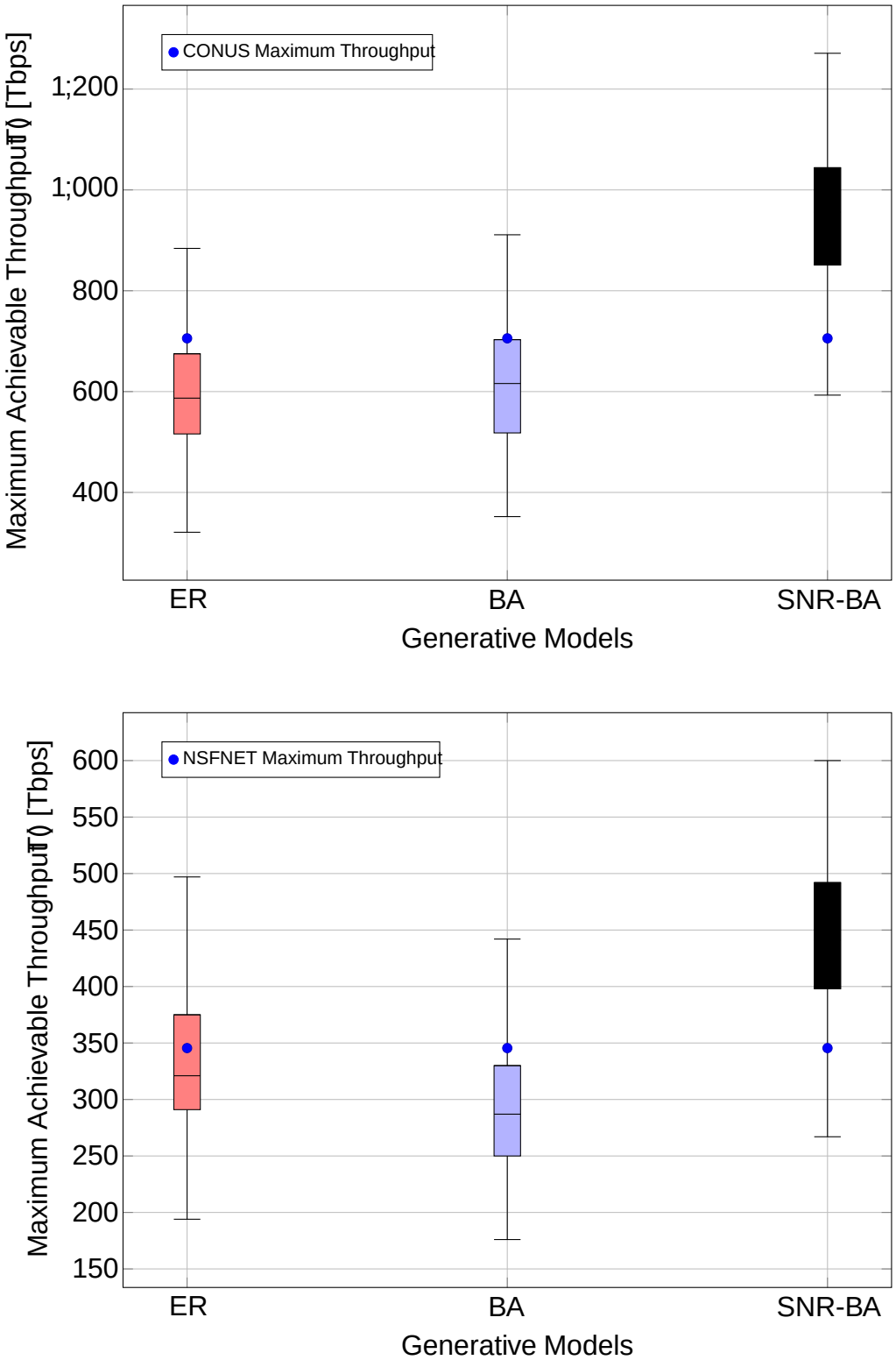


Fig. 3.8: Maximum uniform throughput $T()$ of NSFNET and CONUS based topologies for the graphs generated by ER, BA and SNR-BA generative graph models.

for the CONUS based graphs, the SNR-BA graphs achieve 78% and 71% greater throughput per lightpath than the ER and BA graphs respectively. This is a signi cant difference, solely down to the physical properties of the network.

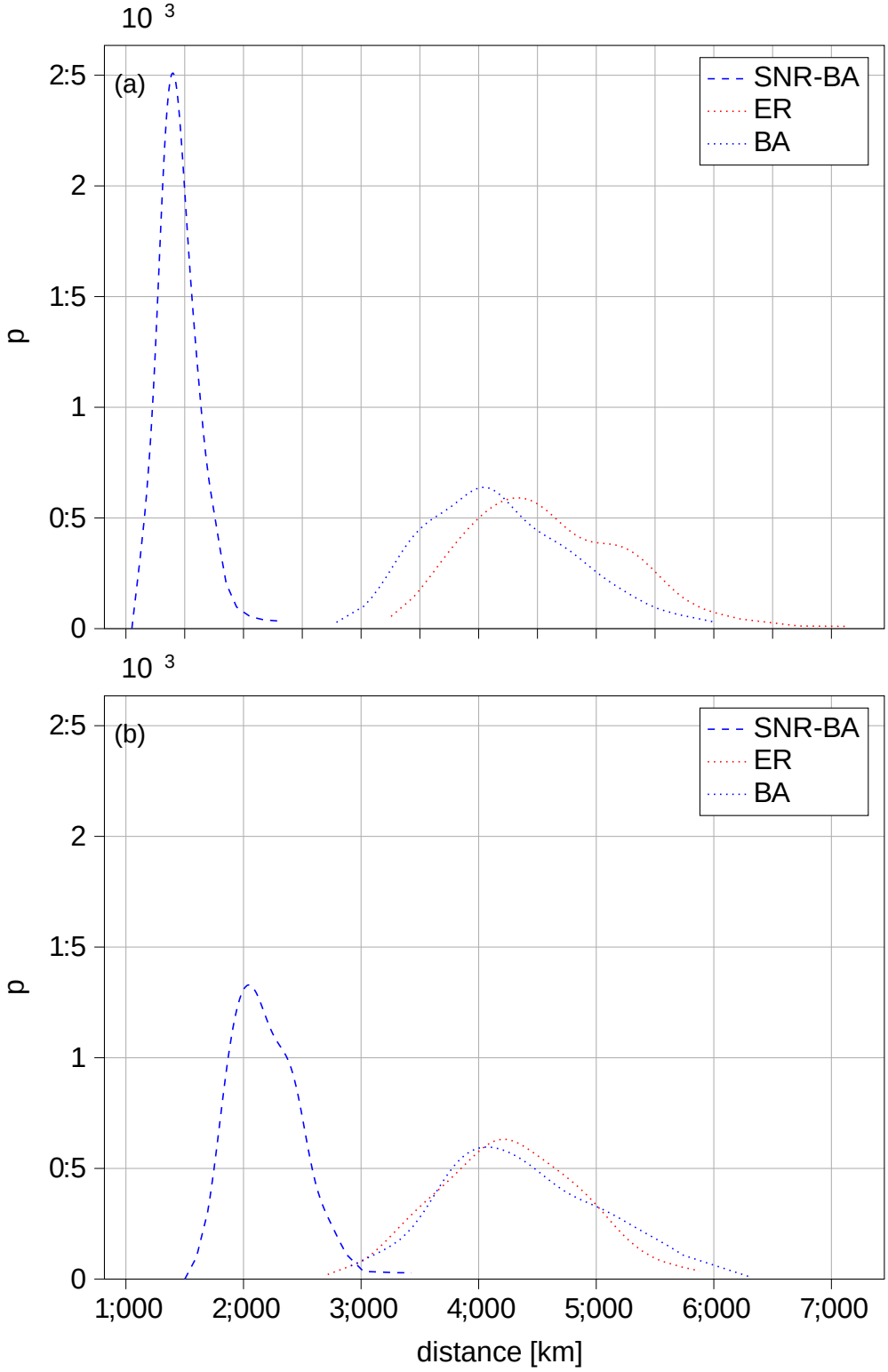
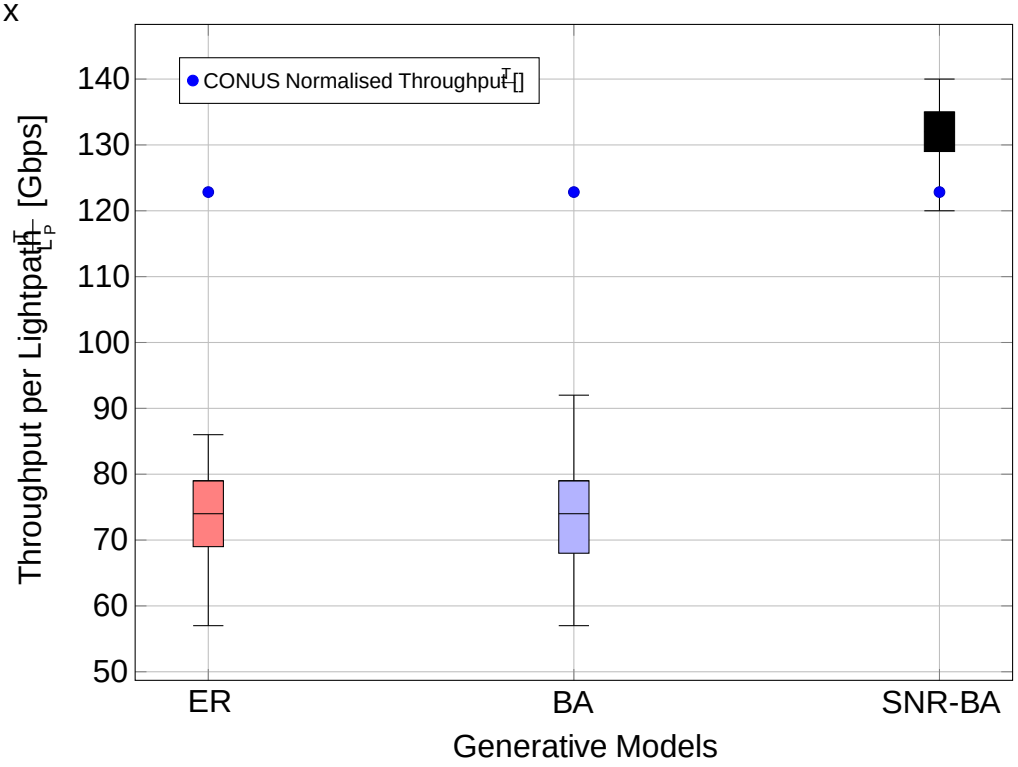


Fig. 3.9: Probability distributions of the average path lengths in each of the solved RWAs for graphs generated by the ER, BA and SNR-BA models, using the node-positions from (a) 30-node CONUS topology and (b) NSFNET.

In summary, the structural advantage of smaller wavelength requirements seen in ER and BA graphs, does not equate to higher throughputs due to the increased path

x



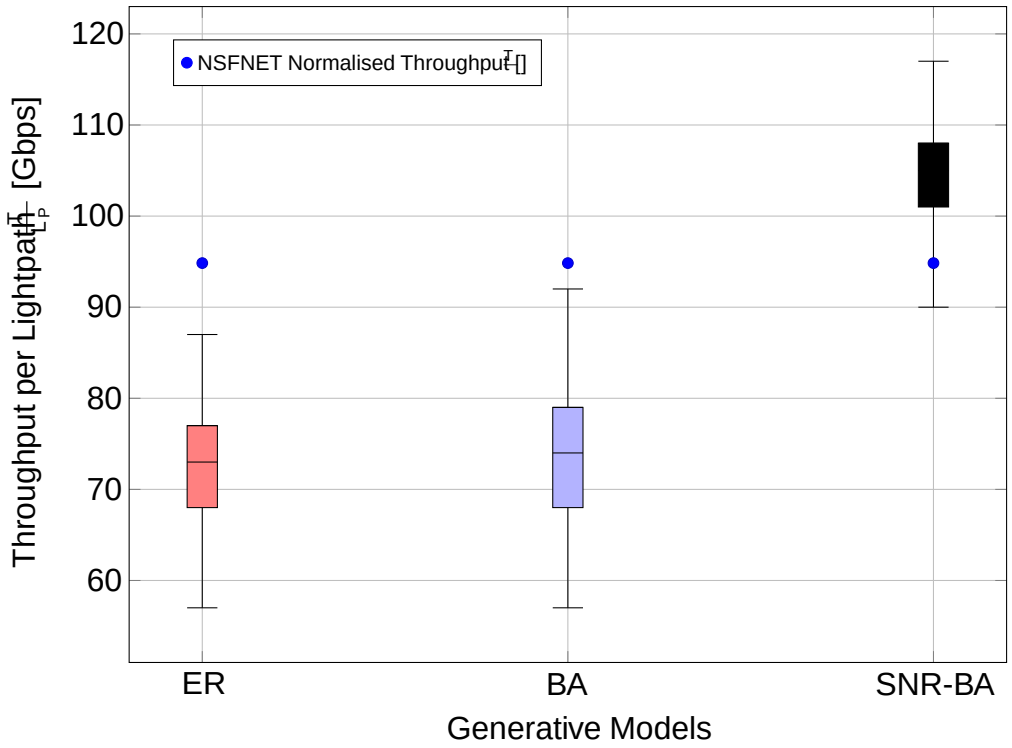


Fig. 3.10: Throughput (T) per lightpath established (L_p) in RWA of NSFNET and CONUS based topologies for the graphs generated by ER, BA and SNR-BA generative graph models.

lengths, and the associated transmission penalties. SNR-BA graphs, however, select shorter edges within the graph generation process and minimise the path lengths and,

Fig. 3.11: Radar plots showing throughput $T(\cdot)$, maximum achievable throughput per lightpath (T/L_P), average lightpath length $P(\cdot)$, number of edges $|E|$, total fibre deployed $L(\cdot_f)$, the averages of the edge length $L(\cdot_e)$.

therefore, help maximise throughput when the distance dominates the achievable throughput in the network. This is summarised in Figure 3.11, where the average values of the edge distances $L(\cdot_e)$, total deployed fibre length $L(\cdot_f)$, number of edges

(jEj), lightpath lengths ($\overline{P}$), throughput per lightpath ($\overline{T}$) and maximum achievable throughput (T) are plotted for each of the generative models and the corresponding network they are based on. For distance and edge values, the scales are reversed so that points on the periphery are better (shorter/smaller), whereas for throughput values the points on the periphery are larger. It is clear that the SNR-BA graphs achieve similar edge numbers, total deployed fibre, edge lengths compared to the original networks, yet lower average lightpath lengths which achieve higher total throughput and throughput per lightpath, than the original and ER/BA networks. Here only graphs that heavily weight physical properties in their generation (SNR-BA) and graphs that ignore them completely (ER/BA) are investigated. To what extent this weighting of distances is important, is an important follow up for the future. By investigating the Waxman and SNR-BA model with varying $_{wax}$/ $_{wax}$ and $_{l}$ values respectively, the sensitivity of these results could be investigated to the degree to which physical properties play a role in the graph generation.

3.6.3 Distance Scaling

Section 3.6.2 showed that smaller diameter networks such as ER and BA networks have a structural advantage in terms of their wavelength requirements. This allows these networks to allocate more lightpaths, as each lightpath on average uses less wavelength resources in the network. It followed, that although SNR-BA networks on average had much larger wavelength requirements (40% and 130% higher in NSFNET and CONUS respectively), they had large diameters and, therefore, are able to allocated fewer lightpaths (between 8-11% fewer). The throughput achievable on these lightpaths however ended up significantly higher than on the ER and BA graphs, on average 44% and 51% higher respectively. This was due to the smaller physical path lengths - on average 63% and 58% shorter than ER and BA graphs respectively - demonstrating how physical properties affect the maximum achievable throughput within networks. This obviously depends on the distance scale, as with larger distances this will only be more true, however with smaller distances the structural advantages should correlate to higher throughput within the network.

By looking at different distance scales of the generated graphs, one can analyse at which point the structural advantages of the ER and BA graphs are beneficial to the total achievable throughput of a network. To do this, the distances of all edges were scaled by some factor $x \in [0.01, 1]$, after which the throughput was recalculated. For each value of x another 200 values of throughput was calculated per generative model. These were then averaged and plotted in Figure 3.12, along with the throughput scaling results of the original CONUS and NSFNET graphs, to understand at which point structural properties benefit total achievable throughput in optical networks.

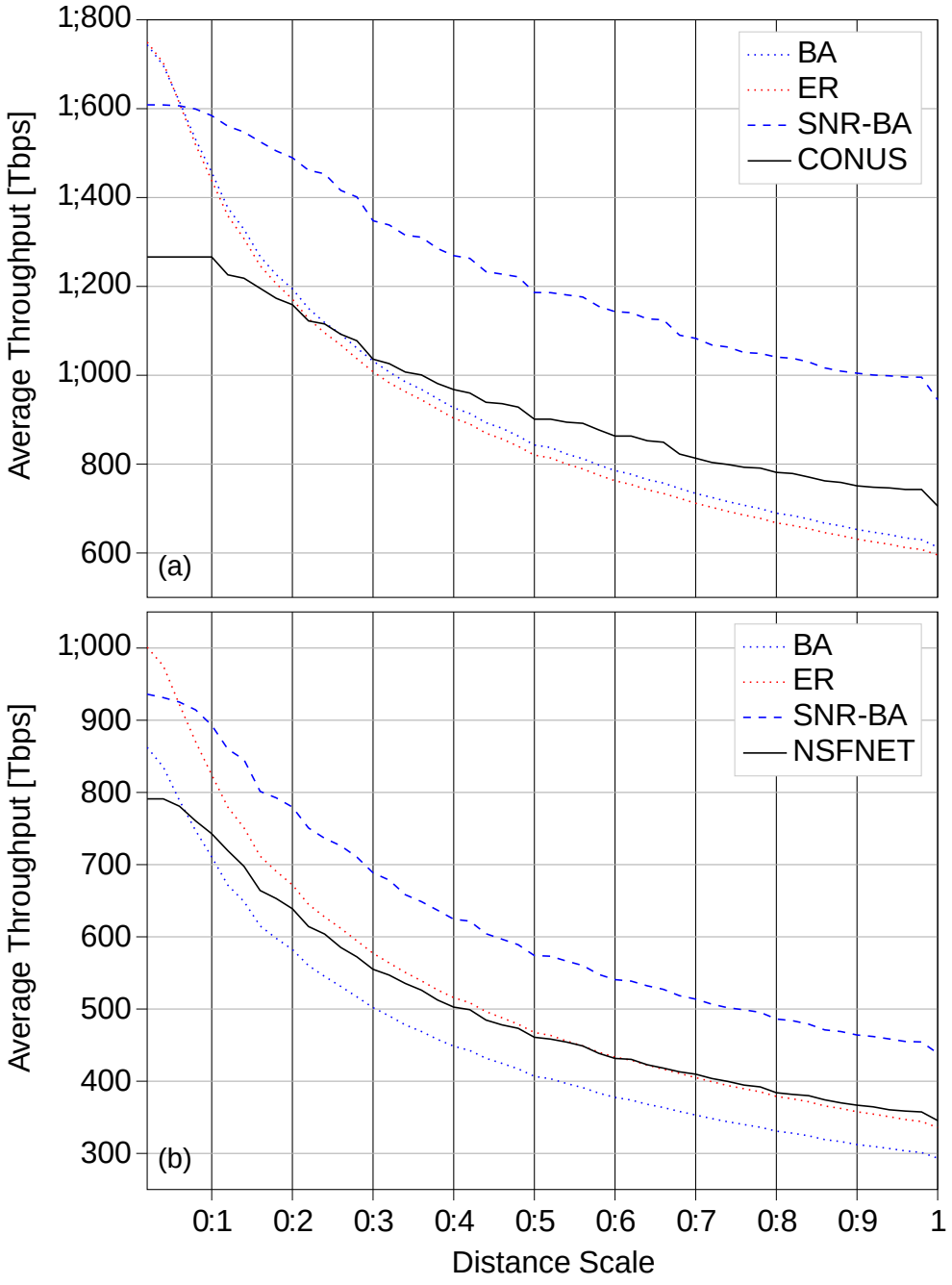


Fig. 3.12: Average maximum achievable throughput (T) calculated for the distance scale x of the graphs generated by the ER, BA and SNR-BA generative graph models based on node-positions taken from (a) 30-node CONUS network and (b) NSFNET network.

In Figure 3.12a one can see that the BA and ER graphs outperform the SNR-BA graphs by about 8% and original CONUS graph by about 38% when looking at distance scales where $x < 0{:}04$. This is reasonable since BA and ER graphs are, on average, able to establish more lightpaths than the SNR-BA and CONUS graphs, however the achievable throughput of these lightpaths suffers due to long edge distances. Once these distances become too small to effectively diminish the throughput of these lightpaths,

the ER/BA start outperforming the SNR-BA and CONUS network. This is down to the structure of the ER/BA graphs as they have more and better connected edges than the SNR-BA graphs and original CONUS network.

Graphs generated for NSFNET node locations were able to achieve the same number of edges due to higher connectivity of the original network. Figure 3.12b shows the average throughput scaling of the networks generated for the three generative models and for the original NSFNET. Here the BA graphs are the worst performing graphs, only outperforming the original NSFNET at smaller distance scales. The ER graphs however outperform the SNR-BA graphs again at scales of $x < 0.1$. This is due to the ER graphs being more uniformly connected across the graphs, leading to 31% lower wavelength requirements and the ability to establish 9% more lightpaths due to structural advantages. This however, comes at the cost of 95% increase in path lengths. However, at smaller scales the distance penalties have a significantly lower or no impact on throughput. As the fibre lengths are scaled via x, the path lengths grow quicker for the ER and BA graphs due to 108% and 111% larger edge lengths for the ER and BA graphs respectively. This in turn reduces their average performance when looking at the original distance scales - causing SNR-BA graphs to on average out perform the ER graphs by 30% and the BA graphs by 49% in the end.

Therefore, when looking at optical core network distance scales, the SNR-BA graphs that use edge probabilities derived from the SNR between nodes, create topologies that perform better, on average, than the ER and BA graphs, due to their shorter path lengths that are more resistant to scaling of distances and nonlinearities. However, BA and ER graphs tend to create graphs with lower wavelength requirements, which create better performing networks for smaller distance scales. This may not be applicable to core optical networks but may be valuable when designing data centre or access networks.

3.7 Summary

This chapter investigated generative graph models for optical networking. Most generative graph models to-date were not suitable for optical core network research, due to omission of physical properties, or the scaling of physical properties that was not accurate for optical networking. Based on observations by previous research that local optical core networks are grid-like in structure, however with the presence of localised well-connected hubs, a new generative graph model, the SNR-BA model, was proposed based on the BA model. The SNR-BA model was compared with four other previously used generative graph models over a dataset of 25 real optical core networks, to validate whether the model was close in structure to that of real optical networks. By comparing the given degree, diameter and spectral properties of these

networks it turns out that the SNR-BA graphs give networks that are structurally closest to those of the real networks, motivating its further use in optical networking research.

Further analysis based on two real optical networks was conducted by generating different structures using ER, BA and SNR-BA graphs, to further understand how the structure and physical properties of optical networks impact the maximum achievable throughput of optical networks. Evaluating these structures for wavelength requirements, it was shown that the structural properties generated by ER, BA graphs lower wavelength requirements on average by 42% and 37% compared to the SNR-BA networks respectively, as they are better connected with smaller diameters. However SNR-BA networks are more grid-like, with larger diameters and therefore generally give larger wavelength requirements, like the CONUS network. When moving onto maximum achievable throughput however it was demonstrated that although the bene cial structural properties seen within the ER and BA graphs from the wavelength requirement metric, translated to being able to assign more lightpaths in total - 11% and 12% more lightpaths in ER and BA CONUS networks respectively - the total achievable throughput of these lightpaths was less - 37% and 35% less in ER and BA CONUS networks respectively - than that of the SNR-BA networks. This demonstrated the impact of physical properties on the achievable throughput of optical networks.

All the networks were scaled by a factor to understand at which scale physical properties do not dominate the performance. ER and BA graphs ended up outperforming the SNR-BA graphs at scales below 10% of those initially tested, demonstrating at which point it is important to consider these within network design.

The SNR-BA model and the importance of using a dataset of networks to study optical networks (demonstrated at the beginning of this Chapter) motivated the development of an open-source dataset for optical network research, named Topology Bench. This dataset includes an exhaustive 105 real optical network and 270,000 SNR-BA networks for optical network research [113].

The next step is to investigate methods for including maximum achievable throughput within physical topology design. To do this, computationally ef cient methods for estimating this property are explored in Chapter 4.

Chapter 4

Estimating Maximum Achievable Throughput

The last chapter investigated what do existing optical core network structures look like and how can we generate these. This was used as a starting point to investigate how the structures in these networks and their resultant physical properties impact performance properties, in particular the maximum achievable throughput of an optical network.

In the 1990s and early 2000s, when designing networks the two most common cost functions aside from network cost, were (i) connectivity (ii) wavelength requirements [114, 115, 18]. Connectivity is a key factor in the reliability and routing performance of a network [116]. The structure of a network physically limits the wavelength requirements of networks [18], also limiting the maximum number of allocatable lightpaths within the network. In fact, there is a strong correlation between (i) and (ii), shown previously by Baroni et al.[18].

Calculating exact wavelength requirements is a NP-complete problem, making it computationally complex [114]. When designing optical networks, there is a large solution search space, for which one needs to evaluate design objectives, such as the wavelength requirements for many networks. To do this exactly is infeasible. Therefore, much research has investigated, how can one include this objective within network optimisation.

Research has generally focused on finding lower-bounds on wavelength requirements and finding ways to compute these efficiently [18, 93, 26]. This lower bound can be found by framing the problem as an optimisation problem, as in Chapter 3, and finding the RWA that minimises total number of wavelengths used. Initially there was a focus on developing heuristics to make this computationally feasible and achieve this lower bound as in [117, 114].

Bayvel and Baroni took a different approach to this problem, by framing a lower bound in terms of a *limiting cut* within the network [18]. The idea being that one can cut the graph into two subgraphs and the links that span across this cut can be used to bound the wavelength requirement in the network. Obviously, there are many cuts in the network that produce different wavelength requirements, however the cut producing the highest wavelength requirement, is the *limiting cut*.

This limiting cut can be formulated as following. Consider that we have a subset of

edges $c_i \in E$, whose elimination causes the network to have two connected components. The set of cuts that separate the network in two is termed C_L. The number of nodes of these subgraphs can be referred as $|O|$ and $|N| - |O|$. Here we assume that we use all-to-all connections (uniform), which dictates that $|O|(|N| - |O|)$ lightpaths need to be established over $|c_i|$ edges. Then a lower bound on wavelength requirements over this cut can be defined as in Eq.(4.1).

$$w_{c_i} = \frac{|O|(|N| - |O|)}{|c_i|} \tag{4.1}$$

Then one can define the limiting cut over all cutsets as in Eq.(4.2).

$$w_r = \max_{c_i \in C_L}(w_{c_i}) = \max_{c_i \in C_L} \frac{|O|(|N| - |O|)}{|c_i|} \tag{4.2}$$

Finding this limiting cut however is still an NP-hard problem and therefore not computationally efficient for large scale network design, however has recently been revisited using heuristics [19].

Therefore, in the 2000s most research focused on understanding wavelength requirements as a function of graph metrics. Fenger characterised the average wavelength requirement based on the number of spanning trees within a network [93]. Châtelain further investigated this problem, by looking at wavelength requirements against other topological metrics, including the algebraic connectivity (λ_1), which is defined as the second smallest eigenvalue - as in section 3.1.3 - of the Laplacian matrix of a graph [26]. This was determined to have the highest correlation to wavelength requirements, which theoretically makes a lot of sense. Baroni already concluded that wavelength requirements depend mainly on the connectivity of a network. Algebraic connectivity can measure how well the network is connected, in terms of how easy it is to disconnect the network. This metric was also investigated to measure resilience of a network [25]. This again is closely related to the wavelength requirement, since lower wavelength requirements, will need to cut more edges to disconnect the graph, making it more resilient. Yuan further looked at this problem and correlated average shortest path length to the wavelength requirement of a network, again related to the limiting cut and algebraic connectivity [89]. Therefore, it is well known which graph parameters to optimise, to minimise the wavelength requirements of networks. However, as seen in Chapter 3, networks with low wavelength requirements, do not necessarily directly translate to high maximum achievable throughput. This was shown to be down to longer edge and path lengths, resulting in worse achievable throughput of routed lightpaths. Demonstrating that the physical properties of networks are an important factor, complicating matters further.

Maximum achievable throughput as defined in this thesis first appeared as a metric of interest in 2016 [118, 76]. Jyothi et al. looked at throughput as a measure of

performance for data centre networks. In data centre networking, cut-based metrics are often used to determine this, like that of sparsest cut and bisection bandwidth [118]. Sparsest cut refers to the cut that has the minimum ratio of capacity to net weight of ows owing over the cut. This can be thought of as an equivalent to the limiting cut in optical networks. Bisection bandwidth however is traf c independent and simply refers to the capacity of the worst case cut that divides the network in two equal sized parts, independent of any traf c ows. Jyothi et al. argued that these cut based metrics are NP-hard to calculate and only provide a loose upper bound for multi-commodity ow problems. This metric inherently is simpler to compute for data-centre networks, since physical properties are not a problem and the routing problem is not edge-disjoint, meaning that the problem can be solved as a multi-commodity ow problem by using a linear programming (LP) formulation. This can be done in polynomial time.

Alvarado et al. [76] on the other hand de ned maximum achievable throughput for optical networks as is used in this PhD research and solved it using the ILP formulation used in section 3.6.2. Within this paper the RWA problem is an edge disjoint problem and results in an NP-hard problem. Ives used this same throughput de nition and generalised the limiting cut initially used in Baroni's work to include maximum achievable throughput [75]. This is strictly an upper worst-case bound and not a tight bound, as the exact value can only be calculated with the exact allocation of channels. This limiting cut results in an upper bound on maximum achievable throughput and can be characterised by

$$(\overline{T_z^B}) = \min_{c_i \in C_L} |c_i| |W| \left(\sum_{z \in Z} \frac{\overline{T_z^B}}{C_z} \right)^{-1} \tag{4.3}$$

where $|W|$ is the number of wavelengths, $\overline{T_z^B}$ is the normalised bandwidth traf c and C_z is the estimated achievable throughput between node-pair z.

One limitation of this approach again is that it is an NP-hard problem to calculate the cut. In addition, the assumption of using shortest path over the cut is limiting compared to the ILP, however as mentioned, it is not a tight upper bound on the throughput.

Vincent suggested the use of a heuristic, FF-kSP, to solve the RWA problem, showing that it achieves high performance compared to the real maximum achievable throughput in many cases - in the best case 1000 networks generated on the BT22 topology achieve the same average maximum achievable throughput compared to the ILP [7]. However, there are still sets of networks for which it does not perform well. For example the DT9TEST topologies showed that over 60% of topologies did not achieve the ILP maximum. Within network design, this would mean that many good solutions would be excluded, even though, they might actually be better solutions. Additionally, the heuristic is not tested on larger networks ($k > 22$), where the

networks generally are even sparser and the RWA problem becomes even more difficult. The heuristic also relies on sequentially loading the network, which requires running many instances of the heuristic to obtain the blocking point of the network, which scales with network size.

More recently, Namyar further investigated throughput as a measure for data-centre networks, motivated by Jyothi's work, however with the goal of finding a tighter upper bound that is computationally efficient [119]. They found that one can bound the throughput of a data-centre network by Eq.(4.4).

$$\Theta(\overline{T_z^C}) \leq \frac{\sum_{u \in K} (R_u - H)}{\sum_{u \in K} \sum_{v \in K \setminus \{u\}} t_{uv} L_{uv}} \tag{4.4}$$

where H is the number of servers, R_u is the number of network facing ports for node u, t_{uv} is the traffic between nodes u and v, and the shortest path length between u and v - L_{uv}. The limitations on this bound are the use of shortest path lengths, L_{uv}, which is not always true. This derives an upper bound for the multi-commodity flow problem, which is a relaxation of the wavelength continuity constraint of the RWA problem.

Despite much prior work, to-date there is no computationally efficient way of including maximum achievable throughput as the design objective for optimising optical networks. Therefore, the goal within the work of this PhD thesis is to further reduce the computational complexity and accuracy of calculating maximum achievable throughput, with the overarching objective to include this as an objective in network design. This next section investigates whether tight computationally efficient upper bounds on maximum achievable throughput of optical networks can be computed and included in the PTD problem.

4.1 Deriving Upper Bounds

The investigation starts with the same starting point as Namyar's work, by investigating the ingress and destination traffic within optical networks. We define $\Theta(\overline{T_z^C}) \sum_{u \in N} \sum_{v \in N \setminus \{u\}} t_{uv}$ as the maximum achievable throughput of that network, where $\Theta(\overline{T_z^C})$ is a scaling factor of the traffic distribution. Maximum achievable throughput here is first investigated in terms of the number of connections that can be assigned, neglecting the physical properties, which can be incorporated via the GN-model. The transit traffic of node u is defined as $X_u^t(\overline{T_z^C})$ and the destination traffic is defined as $X_u^d(\overline{T_z^C})$. The sum of the transit traffic and destination traffic must be less or equal than the total capacity of that node, which is determined using the nodal degree d_u) and number of wavelength channels within the network, $|W|$, resulting in Eq.(4.5).

$$X_u^t(\overline{T_z^C}) + X_u^d(\overline{T_z^C}) \leq d_u \cdot |W| \tag{4.5}$$

$X_u^d(\overline{T_z^C})$ can be expressed in terms of the scaling factor and traffic matrix and substituted in Eq.(4.5)

$$X_u^t(\overline{T_z^C}) + \sum_{v \in N, v \neq u} t_{uv} \quad d_u \quad |W| \tag{4.6}$$

Summing this inequality over each of the nodes $u \in N$ gives us Eq.(4.7).

$$\sum_{u \in N} X_u^t(\overline{T_z^C}) \quad \sum_{u \in N} d_u \quad |W| \quad (\overline{T_z^C}) \sum_{u \in N} \sum_{v \in N, v \neq u} t_{uv} \tag{4.7}$$

One can also define the total amount of transit traffic within the network by looking at each allocated traffic and then looking at the path lengths ($len(k)$), which are used to route this traffic. Using these lengths minus one hop (due to the terminating of the traffic in the last hop), one can determine this total transit traffic $\sum_{u \in N} X_u^t(\overline{T_z^C})$ with Eq.(4.8).

$$\sum_{u \in N} X_u^t(\overline{T_z^C}) = (\overline{T_z^C}) \sum_{u \in N} \sum_{v \in N, v \neq u} t_{uv} \sum_{k \in K_{uv}} (len(k) \quad 1) \tag{4.8}$$

Eq.(4.8) however uses a routing and variable path lengths. One can substitute $len(k)$ with the shortest path length L_{uv} stating that the sum of transit traffic $X_u^t(\overline{T_z^C})$ must be greater than equal to:

$$\sum_{u \in N} X_u^t(\overline{T_z^C}) \quad (\overline{T_z^C}) \sum_{u \in N} \sum_{v \in N, v \neq u} t_{uv}(L_{uv} \quad 1) \tag{4.9}$$

By substituting Eq.(4.9) in Eq.(4.7), one can obtain an upper bound on throughput given by Eq.(4.10).

$$(\overline{T_z^C}) \quad \frac{\sum_{u \in N} d_u \quad |W|}{\sum_{u \in N} \sum_{v \in N, v \neq u} t_{uv} \quad L_{uv}} \tag{4.10}$$

This upper bound, however is not a tight upper bound if $|K_{uv}| > 1$ (the number of paths used for routing). For that one would need a path distribution to calculate a tighter bound, however this is difficult to estimate for optical networks. This is because for maximum achievable throughput, one needs to evaluate the optimal path distribution, not just for the routing problem (maximum flow problem), but also includes the wavelength assignment problem.

4.1.1 Verifying the Throughput Upper Bound

The maximum achievable throughput $(\overline{T_z^C}) \sum_{u \in N} \sum_{v \in N, v \neq u} t_{uv})$ here is given in terms of the number of connections that can be established within the network given a specific distribution of traffic. To calculate this number exactly for optical networks, one needs to formulate the optimisation problem of maximising the number of connections

within the network as an ILP problem, this is different to the formulation in Chapter 3, which maximised the bandwidth demanded between each node-pair.

Maximising the number of lightpaths allows the problem to be speci ed in terms of an integer objective making the problem less computationally complex, than the previous ILP investigated in chapter 3. A decision variable, $_{wkz}$, is used to de ne a lightpath, where W, K_z and Z, are the set of wavelengths, k-shortest paths and node-pairs, respectively. It is de ned in Eq.(4.12), and is constrained to assigning a lightpath, subject to the normalised traf c matrix $\overline{T_z^C}$ and the objective $_{ILP}$, de ned in Eq.(4.13). $_{ILP}$ is a scaling factor of the traf c matrix $\overline{T_z^C}$, which if maximised, maximises the number of lightpaths routed in the optical network via Eq.(4.13). For this work uniform traf c was used, meaning that uniform bandwidth was assumed to be routed between all node-pairs.

Uniform traf c is not a realistic assumption, as traf c in real core networks rarely, if ever, follows this kind of all-to-all traf c distribution. However, it is not well-known what kind of traf c distributions occur, or will occur in the future in optical networks. It used to be thought that population distributions can accurately model this [120], however nowadays with much compute and services located within huge datacentres and huge AI workhouses of 100,000s of GPUs has changed this. All-to-all traf c is one of the most dif cult distributions to accomodate, as equal capacity needs to exist throughout the network to accomodate this, which is dif cult to achieve in core optical networks. Therefore, as a starting point uniform (all-to-all) traf c is used.

The objective is to maximise $_{ILP}$, as summarised in Eq.(4.11). The wavelength continuity and edge-disjoint constraints of paths are de ned in Eq.(4.14), where $I(j \in k)$ refers to whether edge j occurs on path k.

$$\max(\ _{ILP}\ ) \tag{4.11}$$

$$_{wkz} = \begin{cases} 1 & \text{if } (k, w) \text{ is the lightpath assignment for node pair } z \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

$$\sum_{w \in W} \sum_{k \in K_z} {}_{wkz} = d\ _{ILP}\ \overline{T_z^C} e \quad \forall z \in Z \tag{4.13}$$

$$\sum_{z \in Z} \sum_{k \in K} {}_{wkz} I(j \in k) \quad 1 \quad \forall j \in E \quad \forall w \in W \tag{4.14}$$

A dataset of 6,000 networks between 10 and 15 nodes was generated by uniformly randomly scattering the nodes on a grid the size of the north-American continent and then connecting them via the SNR-BA model, described in Chapter 3. Using this ILP

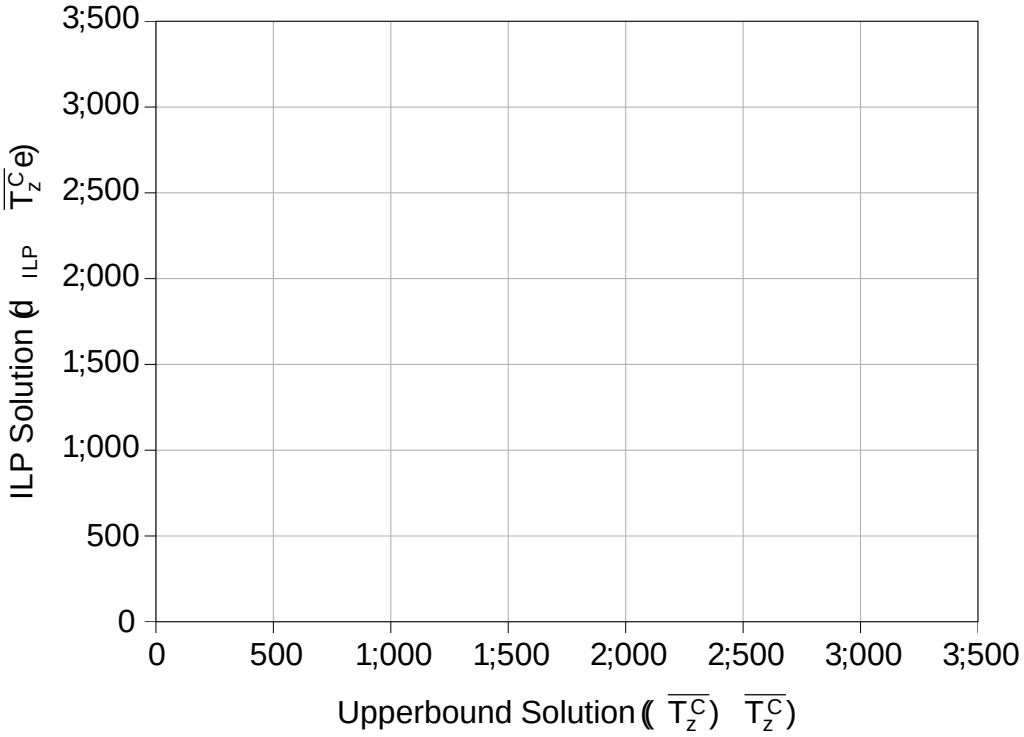


Fig. 4.1: Maximum achievable throughput bound, calculated by summing node capacities versus their true achievable values, calculated via an ILP formulation.

formulation, optimal RWAs were found for each graph in this dataset, using 100 wavelengths and 5 k-shortest paths.

The maximum achievable throughput bound, as de ned in Eq.(4.10), was calculated for all the graphs and plotted against the maximum achievable throughput achieved by the ILP solution in Figure 4.1. Here it is clear that this bound is an upper bound, however not a tight bound, i.e. many bound values are much larger than the ILP solution. This overestimation in maximum achievable throughput by Eq.(4.10), is down to the sum operation over all the nodes in the network. By taking the sum over the whole network, spare capacity in one node can alleviate the congestion of another node, which is not the case when solving the RWA problem in reality.

4.1.2 Tighter Throughput Bounds

The throughput bound de ned in Eq.(4.10) is shown to bound maximum achievable throughput in Figure 4.1. However, due to the huge overestimation of maximum achievable throughput by this bound, it is extremely loose, why is this?

This is because of two assumptions. First is that the use of shortest paths limits the solution to only single path routing, which generally is well known not to be optimal in optical networks (Figure 15 in [7]). In addition to this, taking the sum over all

nodesn assumes that if there is some bottleneck in the network (at some node u), this node can borrow resources wherever they are spare in another part of the network. These assumptions lead to a gross overestimation in maximum achievable throughput of networks. Therefore, this bound is not tight. A tight bound is a bound which cannot be improved upon, i.e. there exists no other that is lower.

Both assumptions made above are addressed to make this bound tighter. Again we can say that we bound the amount of transit traf c $X_u^t(\overline{T_z^C})$ and source traf c $X_u^s(\overline{T_z^C})$ originating from node u, by the maximum amount of resource $d_u \cdot |W|$, as de ned in Eq.(4.15).

$$d_u \cdot |W| \geq X_u^t(\overline{T_z^C}) + X_u^s(\overline{T_z^C}) \tag{4.15}$$

For the following we investigate the sources (s) and destinations (d) nodes within the traf c to investigate the transit traf c over a particular node u in the network. Here the exact path distribution $p(k_{sdk})$ is taken into account - where k_{sdk} is the k-th path between source and destination nodes s and d - which tells us which proportion of paths are used. Therefore, the transit traf c of a node u is de ned as in Eq.(4.16), where $I(u \in k_{sdk})$ is an indicator function, indicating whether node u is in path between source s and destination d over path k.

$$X_u^t(\overline{T_z^C}) = (\overline{T_z^C}) \cdot \sum_{s \in N \setminus u} \sum_{d > s \in N \setminus u} \sum_{k \in K_{sd}} t_{sd} \cdot I(u \in k_{sdk}) \cdot p(k_{sdk}) \tag{4.16}$$

Source traf c can be similarly de ned as in Eq.(4.17).

$$X_u^s(\overline{T_z^C}) = (\overline{T_z^C}) \sum_{d \in N \setminus u} t_{ud} \tag{4.17}$$

Substituting both Eq.(4.16) and Eq.(4.17) in Eq.(4.15) and rearranging for $(\overline{T_z^C})$ gives us the following inequality:

$$\frac{d_u \cdot |W|}{2 \sum_{s \in N \setminus u} \sum_{d > s \in N \setminus u} \sum_{k \in K_{sd}} t_{sd} \cdot I(u \in k_{sdk}) \cdot p(k_{sdk}) + \sum_{d \in N \setminus u} t_{ud}} \geq (\overline{T_z^C}) \tag{4.18}$$

Eq.(4.18) however de nes this quantity for each node in the network, therefore we need to nd the minimum value over all nodes, as this will be the node that causes the bottleneck within the network. This is because if scaling the traf c $\overline{T_z^C}$ any more than this minimum, means that more traf c will be owing over the most loaded node than resources available. Therefore the upper bound $(\overline{T_z^C})$ can be de ned as $_N(\overline{T_z^C})$:

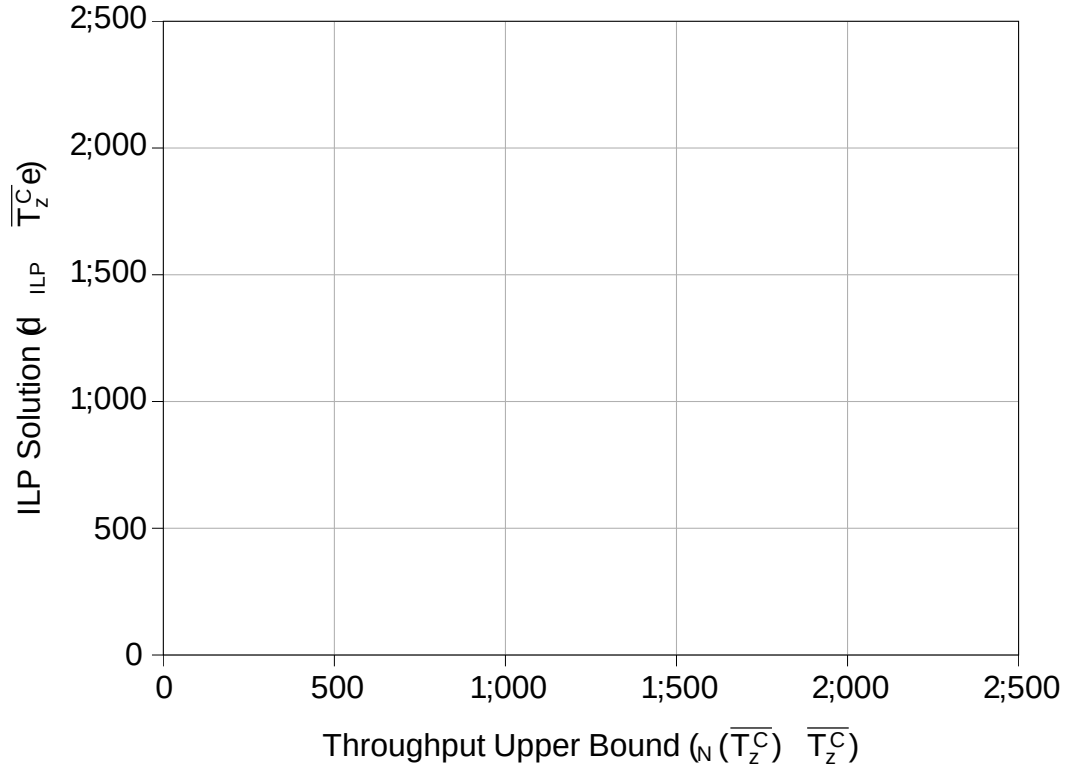


Fig. 4.2: Tighter maximum achievable throughput bound based on node capacities versus their true maximum achievable throughput values.

$$_N(\overline{T_z^C}) \quad \min_{u2N} \frac{d_u \quad jWj}{2 \sum_{s2N\,nu} \sum_{d>s\,2N\,nu} \sum_{k2K_{sd}} t_{sd} \quad I(u\,2\,k_{sdk}) \quad p(k_{sdk}) + \sum_{d2N\,nu} t_{ud}} \tag{4.19}$$

The optimal path distributions, given by the ILP described in section 4.1.1, were extracted and then used to calculate the tighter throughput bound, as de ned in Eq.(4.19). This tighter upper bound is then plotted against the ILP maximum achievable throughput values for the 6,000 sample networks in Figure 4.2. To measure the accuracy of the bound, the coef cient of determination (R^2) is calculated [121]. The coef cient of determination is a measure of how well the variance between two distributions match, normally with values ranging between 0 and 1. A score of 1 would mean a perfect match between the two distributions.

One can observe that the throughput values calculated from Eq.(4.19) are closer to the ILP calculated values than those calculated with Eq.(4.10). The R^2 value between the two sets of maximum achievable throughput plotted in Figure 4.2 is $R^2 = 0{:}89$. This shows that there is a portion of the data that matches well, however not perfectly. In Figure 4.2, 66% of the networks still overestimate the bound by more than 10%.

Fig. 4.3: Demonstration of how only looking at node constraints, fails to fully capture the bottle necks in networks.

Although a huge decrease from the previous bound, which overestimated 69% of the networks by 100%, this still is a large overestimation of maximum achievable throughput. How is this quantity still overestimating maximum achievable throughput, given the optimal path distributions?

There are two assumptions that have been made within this bound: (i) the path distributions k_{sdk} used (ii) node capacities are the bottlenecks. Above (i) was mitigated by using optimal path distributions from the ILP. Therefore, (ii) needs to be further investigated.

A scenario in which the maximum achievable throughput is overestimated is considered. Assume number of wavelengths $W = 4$, and the degree $d_u = 4$ and the traffic flowing in the network is demonstrated as in Figure 4.3. Here one can see that using Eq.(4.18), one gets a value of $\frac{8}{3}$. However, it is clear that the throughput of the network (assuming this is the minimum in the network), is at maximum 2 due to two wavelengths being free on the top two edges and not more. Therefore, its clear to see, that when looking at the whole node, the same phenomena happens in counting resources as we saw initially when summing over the whole network. The summation used over the whole network, as described in Eq.(4.10), resulted in a gross overestimation of maximum achievable throughput. A similar effect is occurring here, where the summation of available resources over the node is used to estimate the

bottle-neck. Again, this results in the overestimation of achievable throughput, as only the top two edges contribute to the bottle-neck, not the bottom two.

Therefore, to improve the bound one can count the resources over the edges rather than the nodes. This is de ned in Eq.(4.20)

$$_{E}(\overline{T_z^C}) \quad \min_{e2E} \frac{P \quad P \quad P \quad jWj}{s2N \quad d>s2N \quad k2K_{sd} \quad t_{sd} \quad I(e2k_{sdk}) \quad p(k_{sdk})} \tag{4.20}$$

The upperbound values for the same 6,000 sample networks were calculated and plotted in Figure 4.4. This new bound overestimates only 0.05% of networks more than 10% of the ILP calculated throughput bound, with $R^2 = 0:99$. This shows that this bound has considerably tightened the estimation of maximum achievable throughput. However, there are both cases where the bound is higher and lower than the ILP calculated case. These cases come from two artifacts within the calculation of the optimal routing and wavelength assignment and the calculation of $_E(\overline{T_z^C})$. Firstly, the ILP was not run to optimality in each case. It was run for 6 hours and the best result recorded. Therefore the optimality gap is not always zero. Secondly, the fractional path probability $p(k_{sdk})$ is a relaxation on the wavelength constraint and does not behave like this in the actual routing and wavelength assignment. Therefore, there are inaccuracies that propagate when taking the fraction over jWj. Therefore, both minor over and under estimations are seen within Figure 4.4.

The problem is that in this case to check the validity of these bounds, the optimal path distributions, extracted from the ILP solutions to the RWA problem, were used and, therefore, these bounds are accurate. These distributions are not easy to calculate and are not computationally ef cient. Can one estimate these path distributions in a computationally ef cient way?

4.1.3 Calculating Path Distributions

One needs to evaluate $p(k_{sdk})$, to nd the throughput bound as de ned in Eq.(4.19), i.e. the distribution of paths used within the routing. Finding the exact distribution of paths used for the maximum achievable throughput, requires the solution of the ILP for the RWA problem, landing us at square one again. To improve the ef ciency of these calculations, one can relax the constraints of the problem, however the question is, how much does this impact the actual upper bound of the network. This idea was summarised best by David Ives [122]:

> is the maximum multicommodity ow calculated using a fully constrained ILP and represents the maximum network throughput. $_{UB}$ $_f$. We could also insert an estimation based on a

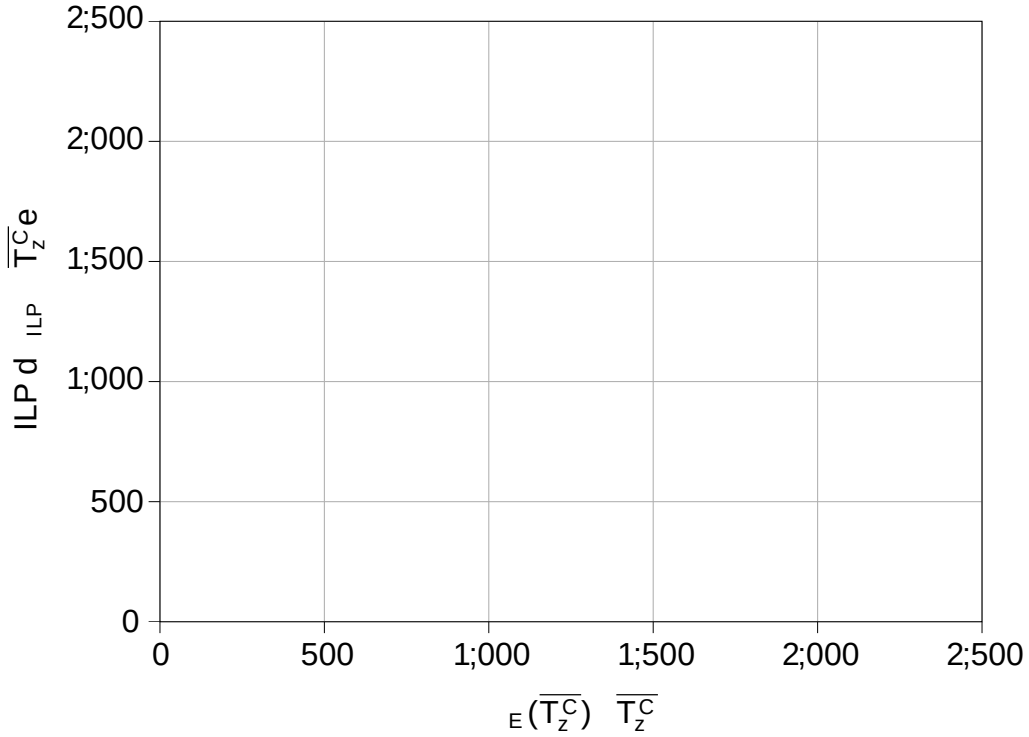


Fig. 4.4: Tighter maximum achievable throughput bound based on edge capacities versus their true maximum achievable throughput values.

wavelength continuity relaxed ILP between and $_{UB}$. The order of these results can be intuitively understood as in moving from the right most $_{f}$ with minimal constraints each move left adds additional constraints until contains all routing, wavelength and integer channel allocation constraints.

$_{UB}$ and $_{f}$ are bounds based on limiting cuts found based on integer and fractional bandwidth assignments respectively [122]. They both ignore wavelength assignments throughout the network, therefore being more computationally ef cient to calculate. The problem obviously is how much accuracy is lost by reducing wavelength constraints on the problem. The problem of nding the path distribution $p(k_{sdk})$ was formulated by the following LP and the results compared with the ILP calculated throughput values.

The optimisation problem can be de ned in the following maxmin problem:

$$\max \min_{u2N} \frac{P}{2} \frac{P}{s2N\,nu} \frac{P}{d>s\,2N\,nu} \frac{P}{k2K_{sd}} \frac{d_u \; jWj}{t_{sd} \; I(u\,2\,k_{sdk}) \;\; p(k_{sdk}) + \frac{P}{d2N\,nu} t_{ud}} \tag{4.21}$$

where $p(k_{sdk})$ is the fraction of paths between source and destination over path k. The variable $p(k_{sdk})$ needs to be further constrained by the constraints in Eq.(4.22)

and Eq.(4.23).

$$\sum_{k \in K_{sd}} p(k_{sdk}) = 1 \quad \forall s \in N \quad d > s \in N \tag{4.22}$$

$$0 \quad p(k_{sdk}) \quad 1 \quad \forall s \in N \quad d > s \in N \quad \forall k \in K_{sd} \tag{4.23}$$

An intermediary variable is de ned to calculate the routing that solves Eq.(4.21) and constrained by Eq.(4.24).

$$2 \sum_{s \in N} \sum_{u} \sum_{d > s \in N} \sum_{u} \sum_{k \in K_{sd}} \frac{d_u \, jWj}{t_{sd} \, I(u \in k_{sdk}) \, p(k_{sdk})} + \sum_{d \in N} \sum_{u} t_{ud} \quad \forall u \in N \tag{4.24}$$

This can be further simpli ed by inspecting the edge capacities as we did before, rather than the node capacities. This results in Eq.(4.25).

$$\sum_{s \in N} \sum_{d > s \in N} \sum_{k \in K_{sd}} \frac{jWj}{t_{sd} \, I(e \in k_{sdk}) \, p(k_{sdk})} \quad \forall e \in E \tag{4.25}$$

By solving this LP by maximising for , gives us a solution for $p(k_{sdk})$ to use within the $_E(\overline{T_z^C})$ bound. Whether this solution gives a $_E(\overline{T_z^C})$ bound that is close to that, which is calculated by the ILP, is investigated next.

For the 6,000 samples the maximum achievable throughput was calculated by solving the LP de ned in Eq.(4.25) and plotted against the ILP calculated value in Figure 4.5. Encouragingly this gives a very similar result to that of Figure 4.4, which used the optimal ILP path distributions. The LP results give a Pearson's correlation and coef cient of determination (R^2) of 0.99, meaning a close to perfect linear correlation score. Although an increase in over estimations above 10% is seen (2% of the networks are overestimated), this is still at a low level (5%).

This shows that actually relaxing the integer constraints of wavelengths within the routing can estimate the large majority (95%) of networks within 10% of their maximum achievable throughputs. This, however, does not include the physical aspects of maximum achievable throughput.

In addition, although this method is computationally ef cient compared to the ILP, solving many (1000s) of LPs for large scale network optimisation still is a dif cult computational problem. Therefore, other methods need to be investigated to do this within ms.

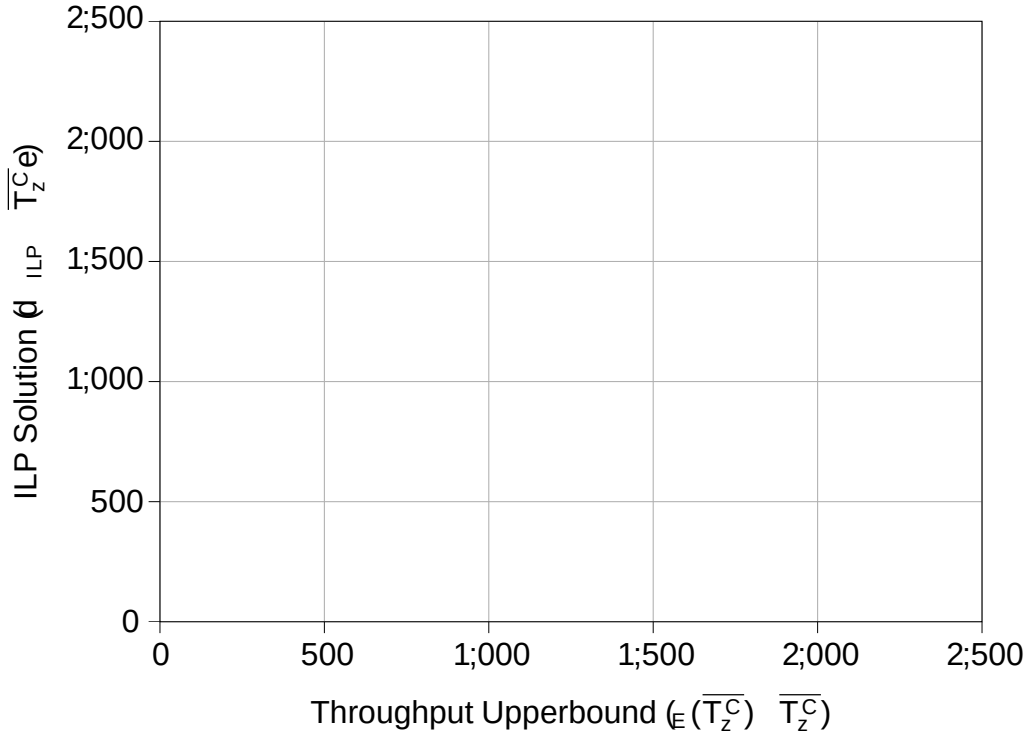


Fig. 4.5: Maximum achievable throughput bound based on edge capacities and LP-found path distributions versus their true maximum achievable throughput values.

4.2 Learning Network Throughput Representations

Throughout this thesis it has been argued that to calculate the maximum achievable throughput of optical networks exactly, or even approximately, is a non-trivial problem. As explained in Chapter 2, this comes from the complexity of solving the optimal routing and wavelength assignment using an ILP formulation, which generally are NP-hard and is not feasible for networks larger than 30 nodes. Therefore, solving the optimisation problems associated with this metric are not necessarily an option when wanting to include it within network design. Therefore, there is still a requirement for a method to evaluate many different graphs (1000s) in very short time (ms).

Therefore, there has been much work in the last two decades on how to approximate these cost function evaluations. Polynomial regression models, linear approximation and deep learning are methods that have been explored for this exact purpose [26, 123]. Most research has focused on data that is structured in a euclidean way (grid structured). Therefore, as reviewed in Chapter 3, for wavelength requirements and robustness of graphs research focused on nding graph parameters correlating them to these objectives [25, 93, 18, 26, 20]. As the structure and physical

properties of these graphs are inherently complex, it is difficult to find single parameters that explain fully the relationship between the structure, physical properties and the maximum achievable throughput. Finding these hand-crafted features requires much research effort and is not an exact science, as it depends on the data that is used to investigate.

Deep learning takes another approach to this problem. It takes raw information from the dataset and creates layers of features. These intermediate features are not human-generated, however are learnt from distributions observed from the dataset [124]. These features can be learnt by back-propagating gradients throughout these modules and adjusting the weights according to some loss function, defined on the learning objective [124]. These weights can then be optimised using a simple procedure such as stochastic gradient descent. In the 1990s it was thought that learning features that could predict anything was impossible. It was thought that stochastic gradient descent would become trapped in poor performing local-minima [124]. Now it is known that this is not actually a problem, as in reality there normally are many saddle points where the gradient is zero, however the performance is at similar levels. Therefore, it is not important where the algorithm gets stuck, as the points are normally equiperformant [124]. With the advent of huge parallelism through fast graphic processing units (GPU) technologies, it was possible by the late 2000s to widely train these networks for useful applications.

These learning methods however operated on Euclidean data. Within optical networking, data is represented as a graph. This graph-structured data holds rich information and dictates many aspects of performance of optical networks, as we have seen. To exploit deep learning in optical networking, initially ANNs were applied by using these hand-crafted features that were hypothesised to explain some of the regression targets [125]. By doing this, there is a loss of information that is encoded in the structure and the physical properties of the graphs, reducing the learning capability of the ANN itself. In addition, this type of approach makes the network rigid towards networks of different sizes and structures.

Research was then concentrated on reducing this pre-processing stage and finding ways of preserving as much as possible of the topological information. This resulted in what is widely known now as graph neural networks (GNN) [126]. The GNN is a supervised learning framework based on learnable functions, i.e. artificial neural networks or recursive neural networks, however adapted to operate naturally over graphs [126]. This ability to operate naturally over graphs preserves the topological information present in the structure of the graph and makes use of raw node and edge features. Therefore, GNNs do not need a pre-processing step and do not need specific graph feature engineering, correlating features to labels. They can learn from raw features and structures within the graph, to understand complex relationships.

The idea here is to investigate surrogate models as was done in [123], however this time investigating GNNs that can use this topological information. The objective being that once a GNN is trained, its inference times are very quick (ms) and they can batch operate over GPUs, making them very computationally ef cient. The next section will review the previous work on GNNs and their operation.

4.2.1 Graph Neural Networks

The advent of graph neural networks is relatively recent. Scarcelli's 2009 "The Graph Neural Network Model" publication laid the foundations of this model, with many others following in its footsteps [124, 127, 128, 129, 130, 131, 132, 133, 134]. The much cited work of Gilmer in 2017 brought together a binding feature of many of these architectures, namely neural message passing [22]. This work focused on predicting properties of molecules, another computationally dif cult problem. Interestingly this also relies on solving Schrödinger's equation. This paper demonstrated the practical learning capacity of graph neural networks, or more speci cally message passing neural networks (MPNN) [22].

Within communications, the application of GNNs or MPNNs is even more recent. The most notable of which is Rusek's work in 2019, which validated that MPNNs could learn relationships that are computationally easy to calculate, namely average delays in queuing networks [23]. He continued this work further to demonstrate this for optimisation within software de ned networks (SDN) [24], where a message passing neural network was adapted to read in paths of a graph and to predict the delay and jitter of that network, to then optimise the performance of the network. The MPNN was able to predict delay on average within 2.2-2.5% and jitter within 6.1-7.8%. The throughput bound investigated in section 4.1.3 for example was able to on average predict within 1.25% of the ILP calculated value. Therefore, the MPNN seems to be able to model some network properties accurately (within 10%). However, within [24], the work speci cally looks at the properties of routing a network and its performance, rather than the edge performance of the network itself, as is the case within the research of this thesis.

The most notable piece of work within communications that looks at an identical problem of edge performance of networks, investigates how to predict performance within data centre networks and was published by Wang in 2021 [135]. Although very close in motivation, these pieces of research were independently thought of and developed separately, only being discovered after publication. They model the throughput and end-to-end latency of data centre networks using a graph neural network. The throughput of data centre networks however is able to be modelled by a LP, as its analogous to the max ow problem [118]. These are simpler problems to

analyse, as the maximum datarate depends on lowest datarate over an edge in a path. However, within optical core networking it depends on the total length of the path, making the problem more difficult, in addition, the integer constraint for wavelengths within the RWA problem makes the problem even more computationally difficult and unpredictable. Although, in section 4.1, an estimation of maximum achievable throughput was derived using an LP, however this ignores the wavelength constraint for routing and therefore the solutions given by it, might not be feasible when including the wavelength constraint. Therefore, given the success of MPNN models for regression tasks on graph-structured data, it was proposed to investigate their usefulness in optical network design.

4.2.2 Message Passing Neural Networks

GNNs are a type of supervised learning framework within geometric deep learning, where message passing neural networks refer to graph neural networks that operate using neural message passing as discussed previously. MPNNs use abstract vectors, referred to as a node or edge hidden state. In this work, for simplicity, the hidden states are confined to nodes, represented as h_n^t, where t represents the message passing iteration. These hidden states, are vectors that hold embeddings for nodes, i.e. for a specific node, they capture structural information from the rest of the graph. The set of node features/edge features are defined as X_N/X_E respectively, and the set of hidden node states as H_N. These networks can be used for either regression or classification tasks. Regression is the task of finding statistical relationships between a dependent and one or more independent variables. Classification on the other hand is the task of determining, given a set of inputs, what category these inputs belong to. In this investigation, regression is focused on.

The MPNN framework centres around three learnable functions: the message function $M_t(h_n^t; h_u^t; e_{nu})$, the update function $U_t(h_n^t; m_n^{t+1})$ and the readout function $R(H_N; X_N)$, where n is a node in the node set N, t is the message passing iteration out of T_{MP} iterations and u is a node in the neighbourhood of n ($u \in N(n)$). T_{MP} is generally chosen to be in the order of either the average shortest path length or diameter of the graph [23]. There have been several different formulations of these functions in the past [129, 128, 130, 131, 132, 133, 136, 127], however, they all follow the same general neural message passing algorithm.

The MPNN - outlined in Algorithm 5 - is made up of three stages: (i) message passing (line 7) (ii) update (line 8) (iii) readout (line 11). In (i) each node in the graph, requests information (messages) from its neighbourhood $N(n)$. This information (messages) is given by feeding in node and edge information into the message function ($M_t(h_n^t; h_u^t; e_{nu})$). To form the nodal message m_n^{t+1} of node n, we have to aggregate

Fig. 4.6: Process of message passing and readout- (Maximum achievable throughput).

Algorithm 5: Message Passing Neural Network Algorithm [23]
Input : G, X_N, X_E
Output : y

```
begin
    for n to N do
        h_n^1  [x_n; 0; 0; :::; 0];
    end
    for t = 1 to T_MP do
        for n 2 N do
            m_n^{t+1}  Σ_{u2N(n)} M_t(h_n^t; h_u^t; e_nu);
            h_n^{t+1}  U_t(h_n^t; m_n^{t+1});
        end
    end
    y  R(H_N; X_N);
end
```

Fig. 4.7: An example of message passing on a 6 node topology.

the information (messages) given by the neighbourhood $\mathcal{N}$. of n. This can be done via different operations, i.e. averaging, sampling or summing. We obtain the message (m_n^{t+1}) of node n, by summing the messages from the neighbourhood $\mathcal{N}$. of n. This is then fed through an update function (line 8) de ned as $U_t(h_n^t, m_n^{t+1})$, which updates the state (h_n^{t+1}) of each node. These two steps are illustrated in the inner block of the diagram shown in Figure 4.6, where one can see that the process is repeated for each node $n \in N$. This procedure iteratively distributes the information of the graph to every node by collecting local messages and using these to update the new hidden vectors. Figure 4.7 demonstrates this process of message passing for the computation of the node state of node 1, on an example 6-node topology. Here the process is shown for two message passing iterations $t = 1, 2$. Working backwards from t=2 and node 1, it can be seen that the neighbours of node 1 are used, which feed their messages into the aggregation that then are updated. Before this, their neighbours do the same for their states. This demonstrates how the information of the graph is distributed across the graph during message passing.

4.2.2.1 Message Function

The message function is used to extract information from both hidden states of neighbouring nodes and adjacent edge features, thus producing messages (line 7 of Algorithm 5). The functions that are used to formulate these messages are generally learnable functions. Within optical networking, the bre lengths of edges signi cantly impact the transmission performance over these edges, therefore it is essential to include this information in the formulation of messages.

$$M_t(h_n^t; h_u^t; e_{nu}) = A(e_{nu}) \quad h_u^t + b(e_{nu}) \tag{4.26}$$

As shown in Eq.(4.26), we use a matrix ANN $A(\cdot)$ and vector valued ANN $b(\cdot)$ to extract features from the edge feature vectors e_{nu} in the constructed messages. These messages are then aggregated via a sum operation, as seen in line 7 of Algorithm 5 and the inner block of Figure 4.6, to give the future message of node n, m_n^{t+1}. The sum operator is chosen due to it being permutation agnostic and for its simplicity.

4.2.2.2 Update Function

The update function is used to take the information (messages) aggregated from the neighbourhood of node n and learn a new hidden state vector h_n^{t+1} to incorporate this new information in the state (line 8). As this is a sequential process, we use a recurrent neural network (RNN) architecture, which can take previous states into account and learn how much of the previous states to use in the next state, whilst producing new abstract representations of the data. RNNs have been shown to struggle with vanishing gradients during training, therefore a gated recurrent unit was used (GRU). GRUs use reset and update gates to learn how much of the state and input to use or forget depending on the target. They exhibit improved performance over standard RNNs, however have been shown to have reasonable computational complexity [137].

$$U_t(h_n^t; m_n^{t+1}) = GRU(h_n^t; m_n^{t+1}) \tag{4.27}$$

The current state, h_n^t, and the aggregated messages, m_n^{t+1}, are fed in, to produce an updated representation, h_n^{t+1}.

4.2.2.3 Readout Function

The final graph-level aggregation of the states and features is carried out via the readout function (line 11). The aim is to aggregate all the relevant inter-dependencies between the nodes, via the hidden states h_n^T, and represent it as a single vector, on which one can regress to the target outputs (throughput, in our case). The readout function is shown in Eq.(4.28).

$$R(H_N; X_N) = b\left(\sum_n [i(h_n^T; x_n)] \quad j(h_n^T)\right) \tag{4.28}$$

An attention mechanism was used to learn which parts of the hidden vectors, h_n^T, are important for the prediction of the target (in our case maximum achievable throughput). The attention mechanism learns weights in the range $[0, 1]$ that are used to weight a vector, with the goal of learning which parts of the vector are

Fig. 4.8: Data generation process for the maximum achievable throughput labels - SL- sequential loading, T - Maximum achievable throughput.

important for the learning task. This is achieved, by feeding the concatenated hidden vector, h_n^T, and node features, x_n, through an ANN and passing the output through a sigmoid function σ, which normalises the output between $[0, 1]$. Using an element-wise multiplication (Hadarmard product), this vector acts as attention scores to the original hidden vector, h_n^T which is fed through another ANN, Ψ [23]. Summing these operations over all nodes gives us the final vector used for the regression layer. The regression layer, ϕ, consists of a single ANN layer, which reduces the output to a scalar value.

Having defined the MPNN model architecture, it now needs to be trained on graph labels, i.e. maximum achievable throughput, to predict this on all unseen graphs. The next section details the methodology for generating the training and testing datasets.

4.2.3 Training Dataset Generation

A single generative graph model was chosen to create the graph structures for training and testing. These were created via the SNR-BA model introduced in section 3.2. In this generative graph model, distances between nodes are incorporated in the process of edge selection, creating localised hubs within graphs. Nodes are chosen uniformly over a grid, representing the size of the north-American continent, resulting in unique node locations for each graph and therefore giving a greater range of graph structures and physical properties than in the work presented in Chapter 3, where the same real network node locations were used in all generated graphs. The SNR-BA generative graph model then uses these unique node locations to generate the graphs. Inter-node distances were constrained to be at least 100 km.

Three separate datasets were generated to demonstrate the model operation on different size topologies: (i) alpha set with $10 \leq |N| \leq 15$ (75000 graphs), beta set with $25 \leq |N| \leq 45$ (95000 graphs) and gamma set with $55 \leq |N| \leq 100$ (75000 graphs). The nodes increased by 1 and 5 for the alpha and beta, gamma datasets respectively, giving 6, 5 and 10 different node scales for the alpha, beta and gamma datasets respectively. To make sure that the model learns performance trends over a variety of edge numbers, edge numbers were chosen by adding an empirically determined percentage of the nodes for a given graph, multiple times, as seen in Eq.(4.29). Here e_n was chosen to be 0.2, typical of the relatively sparse core networks and i varied between 1 and 10 - so that the graph, empirically, has approximately 20% more edges than nodes, as the sparsest core networks have about 20% more edges than nodes, we have used this as the minimum value [8].

$$|E| = |N| + e_n \quad |N| \; i \tag{4.29}$$

An optimal RWA that maximises the number of allocated lightpaths (simply referred to as the optimal RWA) is required to generate the training label for each network considered. To find these optimal RWAs, an ILP formulation from section 4.1.1 was used for $|N| < 25$, and for $|N| = 25 \quad 100$, a heuristic, FF-kSP was used.

Again uniform traffic was used as a starting point here. However, the uniform traffic used was in terms of bandwidth requested. This means that the uniform matrix was formulated in terms of bitrate requested rather than number of lightpaths. The reason for doing this, was that the final number of setup lightpaths depends on the physical properties of each of the networks. Due to the random scattering of the nodes within the network, this meant that each network had unique physical properties that differed slightly from network to network.

A simple transformation is used to get the traffic distribution, in terms of lightpaths requested ($\overline{T_z^C}$) from the bandwidth requested traffic distribution ($\overline{T_z^B}$).

$$\overline{T_z^C} = \frac{\overline{T_z^B}}{2B_{CH}\log_2(1 + SNR_z)}$$

where SNR_z is an estimation of the worst-case SNR between node-pair z. This gave a variation in traffic distribution for each of the networks used for training.

4.2.4 Training

Using the datasets created as described in section 4.2.3, the three MPNN models were trained to predict the maximum achievable throughput, The node features $x_v()$,

chosen to be the degree of each node (d_n) and its normalised traffic:

$$x_n = \left[d_n ; \sum_{u \in N} \frac{1}{T^c_{(n,u)}} \right]$$

were used to initialise the node hidden vectors (h_n^1). As maximum achievable throughput depends on the physical properties of the fibre links, the overall transmission quality metric is a critical feature. The number of lightpaths carried over an edge or set of edges, as well as the length of paths taken, influences the overall system performance via the achievable signal-to-noise ratio in the nonlinear optical regime. Throughout testing it was seen that the inverse of the SNR (NSR) was a better feature to use, as it is an additive quantity over a set of edges. When calculating the SNR of a path, the NSR is summed and then inversed over individual edges. Therefore, when summing the NSR it has a direct relationship to the SNR of paths between nodes. In an ideal case, optimal launch powers would be found for each wavelength in the network, however due to the computational complexity associated with carrying this out for many networks (240,000 topologies), the assumption of uniform launch powers, i.e. equal input power/lightpath for all lightpaths, was used, similar to other work [7].

A hidden vector size of 16, with 8 message passing rounds was used for all model training, similar to [135, 24]. Larger hidden vector sizes did not seem to provide more accuracy, however added higher computational complexity. The message passing rounds were chosen to work well over the whole spectrum of node scales, where $T = 8$ gave good performance for all. Overfitting - the phenomena when a model fits the training data too closely, with deteriorating performance on unseen test data - was initially a problem. Therefore, to combat this, dropout was utilised - a regularisation strategy where randomly nodes within neural networks are 'switched off' to enhance the generalisation capability of a neural network. A dropout rate of 0.65 with L2 regularisation at a rate of 0.03 was used, where higher values started reducing the accuracy of the model. The learning rate was initialised with a value of 0.001 and decayed exponentially using 10,000 steps at a rate of 0.95. The Adam optimisation algorithm was used to train all models, with the graphs in batches of 50 and a single fully-connected regression layer consisting of 256 neurons was used for the final regression output, where larger values did not provide further accuracy without overfitting. To monitor whether the model was tending to overfit, a validation set of 500 graphs was used to evaluate the performance at each epoch. The training was conducted on a Nvidia V100 16GB GPU, with training generally taking around 72 hours, covering 2000 epochs.

Having trained the three separate models on the different datasets, we used three

separate test sets to gauge the performance of the trained inference models. These were generated identically to the methodology laid out in section 4.2.3, however were unseen (not used for training) by the model. The following section analyses the accuracy, computation time and generalisation capabilities of the MPNN model using these test sets.

4.2.5 Maximum Achievable Throughput

There are three aspects we evaluate to understand how well the MPNN model performs in comparison to other methodologies for estimating the maximum achievable throughput of an optical network: (i) model accuracy (ii) time complexity (iii) generalisation capability to unseen graphs.

, The three test sets of 6000, 5000 and 10,000 graphs, were used to evaluate the accuracy of the alpha, beta and gamma models, respectively. Here 1000 graphs for each node scale were generated. Labels for the graphs in the alpha test set were generated via the ILP formulation, whilst for the beta and gamma test sets, FF-kSP was used. Again, uniform traf c was generated and used for all these results. To measure the accuracy of model predictions, the coef cient of determination (R^2) was used, which measures how much of the variance in the data is correctly predicted by the model, by comparing the prediction variance to that of the original data [121]. Here we de ne R^2 as

$$R^2 = 1 - \frac{\sum_n (y_n - p_n)^2}{\sum_n (y_n - \bar{y})^2} \tag{4.30}$$

where y_n refers to the true label and p_n the predicted label and $\bar{y}$ the mean.

The Pearson's correlation coef cient was used as an alternative metric and determines how linearly the labels and predictions are related, with a valu e of 1 signifying an ideal linear correlation. This coef cient is important in the context of surrogate models or cost functions, as a model might be inaccurate (R^2), however have a high linear correlation (high), which means that although inaccurate it can predict the relative performance of a network well - vital for optimisation.

The label that was evaluated is that of maximum achievable throughput (T_{put}) for each of the 6000 graphs in the alpha test set, the performance was evaluated via ILP, FF-kSP, kSP-FF and MPNN and plotted in Figure 4.9(a). It can be seen that the kSP-FF and FF-kSP heuristics underperform compared to the ILP, giving lower R^2 values in both, although kSP-FF is worse as it has even lower values for R^2 and , as seen in Table 4.1. This signi es poor predictive accuracy of the actual labels (T_{put}). The performance of kSP-FF is worse than that of FF-kSP, as it priotises shorter paths over optimising wavelength selection, this is an expected result seen in [7]. FF-kSP generally spreads the usage more evenly across all network links compared to kSP-FF. This spreading of resources uses slightly more spectrum, however achieves much less

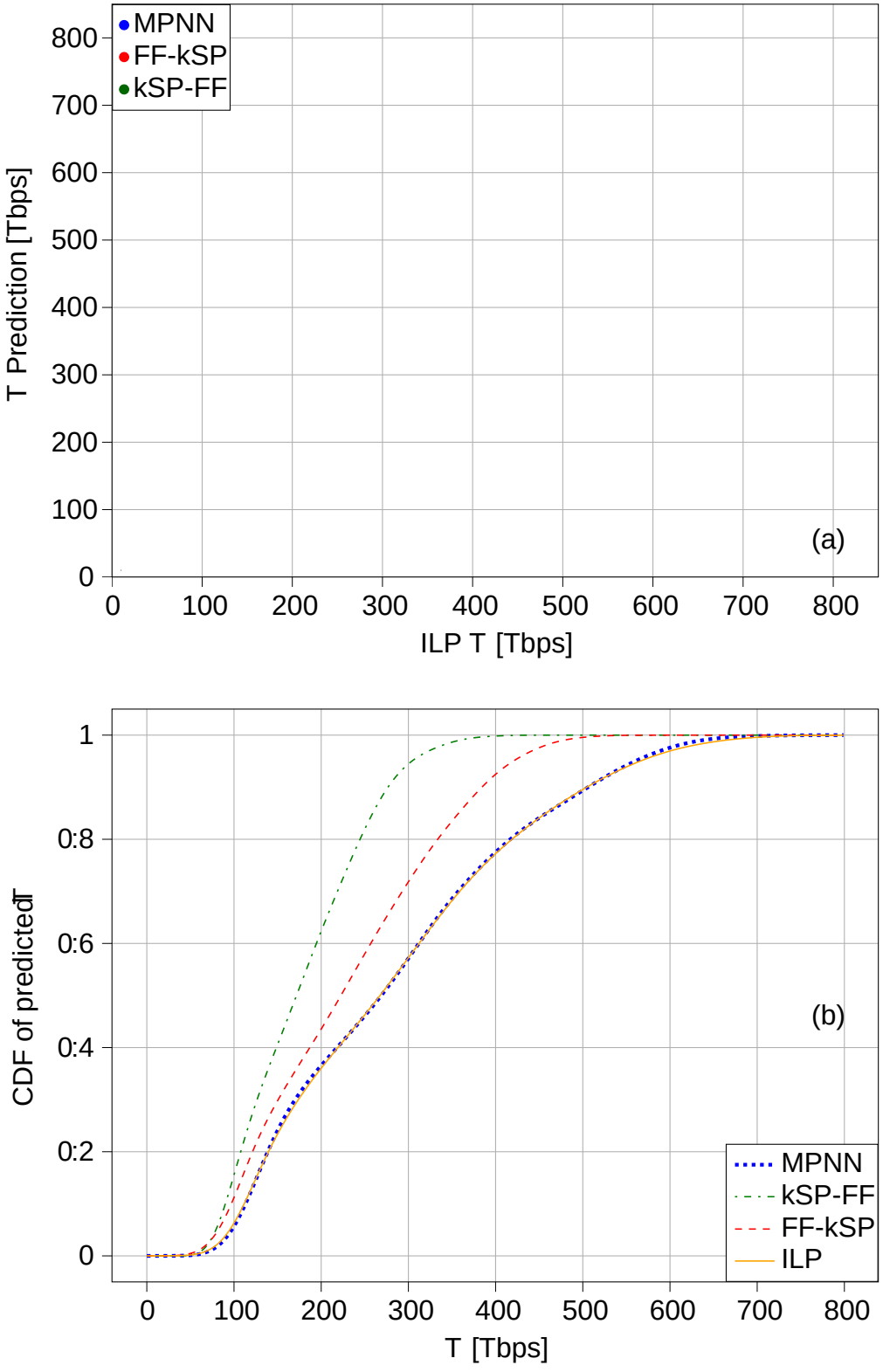


Fig. 4.9: (a) Throughput prediction for FF-kSP, kSP-FF and the MPNN, versus the ILP optimal value for $10 \leq N \leq 15$. (b) The cumulative distribution function (CDF) of the throughput distributions given by FF-kSP, kSP-FF, MPNN and ILP for $10 \leq N \leq 15$.

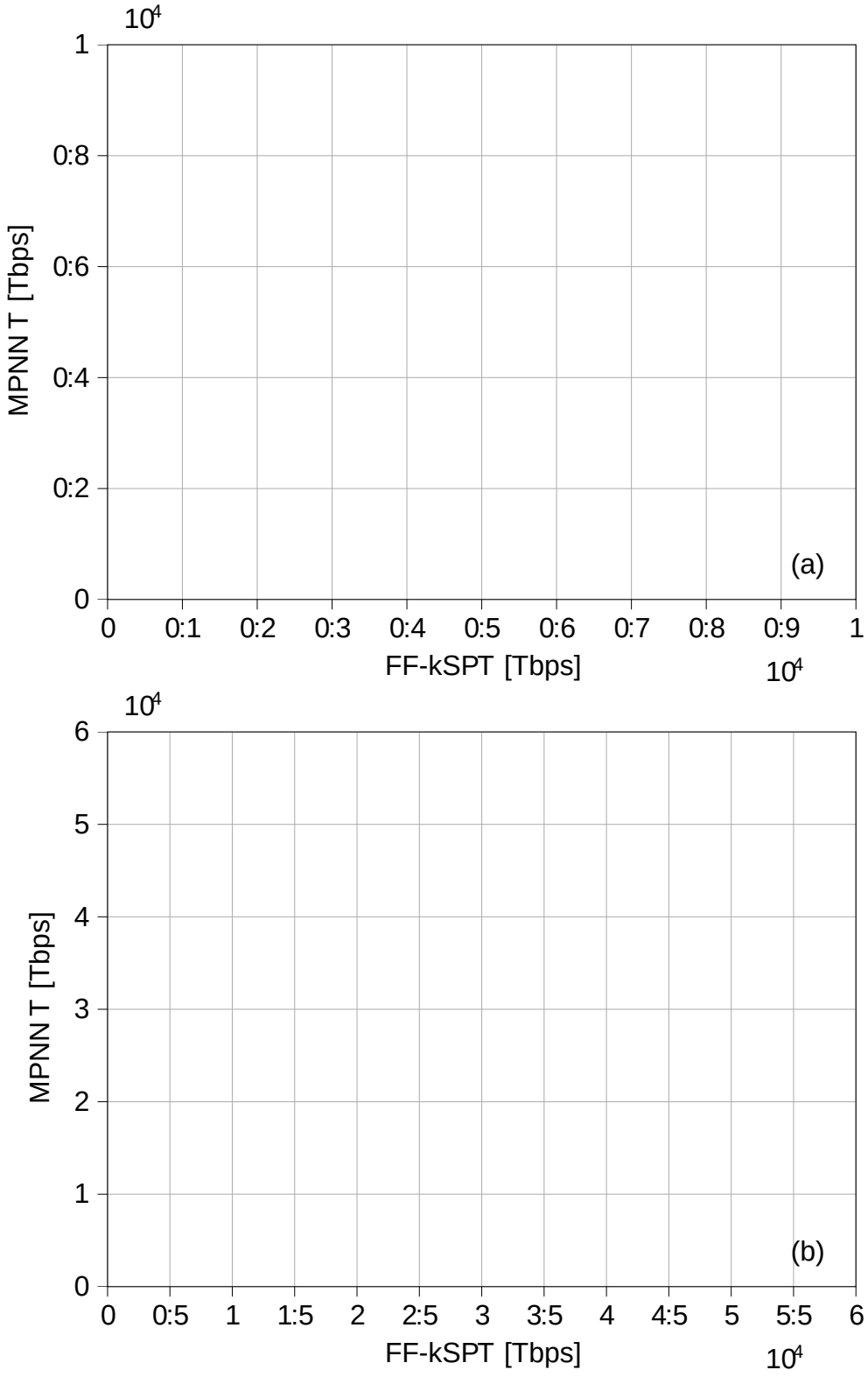


Fig. 4.10: (a) Throughput prediction of the MPNN versus the FF-kSP prediction for 25 j N j 45. (b) Throughput prediction using MPNN versus the FF-kSP prediction for 55 j N j 100.

Node Scale	Method	R^2	
$10 \leq \lvert N \rvert \leq 15$	MPNN	0.951	0.975
$10 \leq \lvert N \rvert \leq 15$	FF-kSP	0.740	0.969
$10 \leq \lvert N \rvert \leq 15$	kSP-FF	0.102	0.901
$10 \leq \lvert N \rvert \leq 15$	ElasticNet	0.763	0.873
$10 \leq \lvert N \rvert \leq 15$	ANN	0.768	0.908
$25 \leq \lvert N \rvert \leq 45$	MPNN	0.973	0.986
$55 \leq \lvert N \rvert \leq 100$	MPNN	0.948	0.974

Table 4.1: Accuracy of the MPNN model and other capacity estimation methods, measured by the coefficient of determination (R^2) and the Pearson's correlation coefficient ( ).

network congestion. However, the MPNN has learnt on a variety of graphs of these sizes and can accurately predict the throughput trend here, as indicated by its R^2 high value of 0.951. On average the MPNN is able to determine the maximum achievable throughput of a network within 8% of the ILP value. The MPNN evaluates 68% of networks within 10% of the ILP calculated maximum achievable throughput. This compares to an average deviation to the ILP calculated maximum achievable throughput of 17% and 33% for FF-kSP and kSP-FF respectively. Only 33% and 8% of the networks were evaluated within 10% of the ILP maximum achievable throughput for both FF-kSP and kSP-FF respectively. Furthermore, one can see the difference in throughput distributions, where the cumulative distribution function (CDF) of the different methodologies are plotted in Figure 4.9(b). Here it is empirically clear that the kSP-FF and FF-kSP heuristics give different distributions compared to the ILP. This can be measured by the KS-2-sample test, as was used in Chapter 3. This measures an absolute largest difference between two CDFs. The MPNN however, is able to replicate the original distribution of the ILP well. This is confirmed by a maximum absolute distance of 0.02, compared to 0.39 and 0.15 for both kSP-FF and FF-kSP. Therefore the MPNN can predict the maximum achievable throughput of the networks better. In addition to the MPNN, a linear and nonlinear ML regression method were evaluated to compare other ML frameworks. Here the degree variance, connectivity, algebraic connectivity, communicability distance and communicability traffic index were used as input features for the graph. ElasticNet and the artificial neural network (ANN) were trained and tested with the same data, both scoring lower with R^2 values of 0.763 and 0.768, respectively.

Another metric evaluated was the Pearson's correlation coefficient ( ). As for the optimisation, there is significant linear correlation between the real performance and the predicted performance properties. It can be seen that the heuristics generally perform well, and the FF-kSP has a high linear correlation ($= (0.969)$ between the estimated throughput values and those calculated via the ILP, as seen in Table 4.1. The MPNN,

has a similar correlation, with = 0:975, meaning it predicts the relative throughput performance of networks well. ElasticNet and the ANN both score values lower than the MPNN and FF-kSP. The high linear correlation of FF-kSP compared to the ILP, makes it a good candidate to evaluate the maximum achievable throughput for the larger graphs, even though the real maximum achievable throughput might be larger.

To further evaluate the performance of the model, the MPNN was applied to the beta and gamma testing datasets that included larger graphs. Here the labels were generated via the FF-kSP heuristic, as the ILP is not able to scale to these larger topologies. The MPNN was used to infer the maximum achievable throughput of the graphs in the respective test sets and plotted in gures 4.10(a) and (b). It is clear that for the beta model, the MPNN performs better than the alpha and gamma models, with a value of 0.973. The beta model was able to predict the calculated value of maximum achievable throughput of about 80% of networks within 10%, with an average deviation from the calculated value of about 6%. This improved performance compared to the alpha set, is because the beta training set had the most training graphs per node scale. This was possible for the beta model, as the heuristic runs faster on the smaller beta graphs than the larger gamma set graphs, and also faster than the ILP. This can be seen in gures 4.11 (a) and (b). The gamma model achieves a high of 0.948, similar to the alpha model. On average it predicts the throughput of a network within 9% of the FF-kSP value and 64% of the networks within 10% of the FF-kSP value. Both the beta and gamma models achieved high linear correlation values ( ) of 0.986 and 0.974 respectively.

One can conclude that heuristics consistently underperform in estimating the throughput in optical networks compared to the ILP solutions, with on average a 17% and 33% deviation from the ILP solutions. However, FF-kSP is able to predict the trend (linear correlation - 0:96) well, making it suitable for optimisation tasks and training data generation and in addition, for many samples it seems to get close to the ILP solution. The MPNN on the other hand is able to accurately predict throughput values of unseen data, based on similar structure and sizes seen during training. It also has a high linear correlation (0:97) between the labels and its own predictions, making it an excellent candidate for modelling the throughput of these optical networks. Therefore, the MPNN is able to predict maximum achievable throughput accurately, on average predicting 70% of networks within 10% of the throughput label. Although it does not provide a solution, for example for the RWA, of how to reach this maximum achievable throughput. The model could potentially be adjusted to learn the RWA, this is outside the scope of the current work and is left for future research. However, given the extensive training and data needs of this model, what is the real bene t of using it in place of heuristics? This question is addressed in the following section.

4.2.6 Computational Time Comparison

The key advantage of using machine learning to model relationships in data, is the very short inference times compared to ILP and, even, heuristic methods.

The ILP has a worst-case computational complexity as described in Eq.(4.31), where D_C is the number of demands or connection requests ($D_C = \sum_{z \in Z} d\overline{T_z^C}$ ILP e), $|E|$ the number of edges and $|W|$ the number of wavelengths used [138].

$$O(2^{D_C |E| |W|}) \tag{4.31}$$

For the heuristic algorithms used, the complexity generally scales as in Eq.(4.32) [139], but must be modified to include the term R_{SL} - number of sequential loading iterations as the heuristic needs to run many times before finding the optimum RWA:

$$O(R_{SL} |K| |N|^3 (E + |N| \log(|N|))) \tag{4.32}$$

The advantage of the MPNN, thus, is that it can directly evaluate the network properties, learnt from previous data. For the inference of the MPNN, the complexity is defined in Eq.(4.33) [140], where T_{MP} denotes the number of message passing rounds, defined in section 4.2.2 and d_h is the length of hidden dimensional vector used, 16 in our case.

$$O(T \quad |N|^2 d_h^2) \tag{4.33}$$

Comparing Eq.(4.31), Eq.(4.32) and Eq.(4.33), it can be seen that the MPNN scales the best, computationally, with number of nodes. To quantify the computational benefits of using MPNNs to model optical networks, graphs with nodes varying from 10 to 20, were used to evaluate the ILP, FF-kSP, kSP-FF and MPNN time performance. Here the node size has been expanded to 20 nodes for the test set to see the computational time scaling better with smaller graphs. This, however, was not possible in the case of the training set, since the ILP computation is too complex. For each graph, the respective methodologies were used to calculate the maximum achievable throughput and their computation times measured and plotted in Figure 4.11(a). The reduction in computation time through using MPNN model can be clearly seen, where it takes approximately 10s of ms, compared to 10s, 100s and 1000s of seconds for kSP-FF, FF-kSP and ILP respectively. The same trend, between FF-kSP and the MPNN, was analysed for the larger networks using the beta and gamma test sets, with the results shown in Figure 4.11(b). The MPNN again shows minimum computation time increase for these node ranges, compared to the heuristic method.

It can be concluded that, the MPNN model can be seen as an accurate and fast

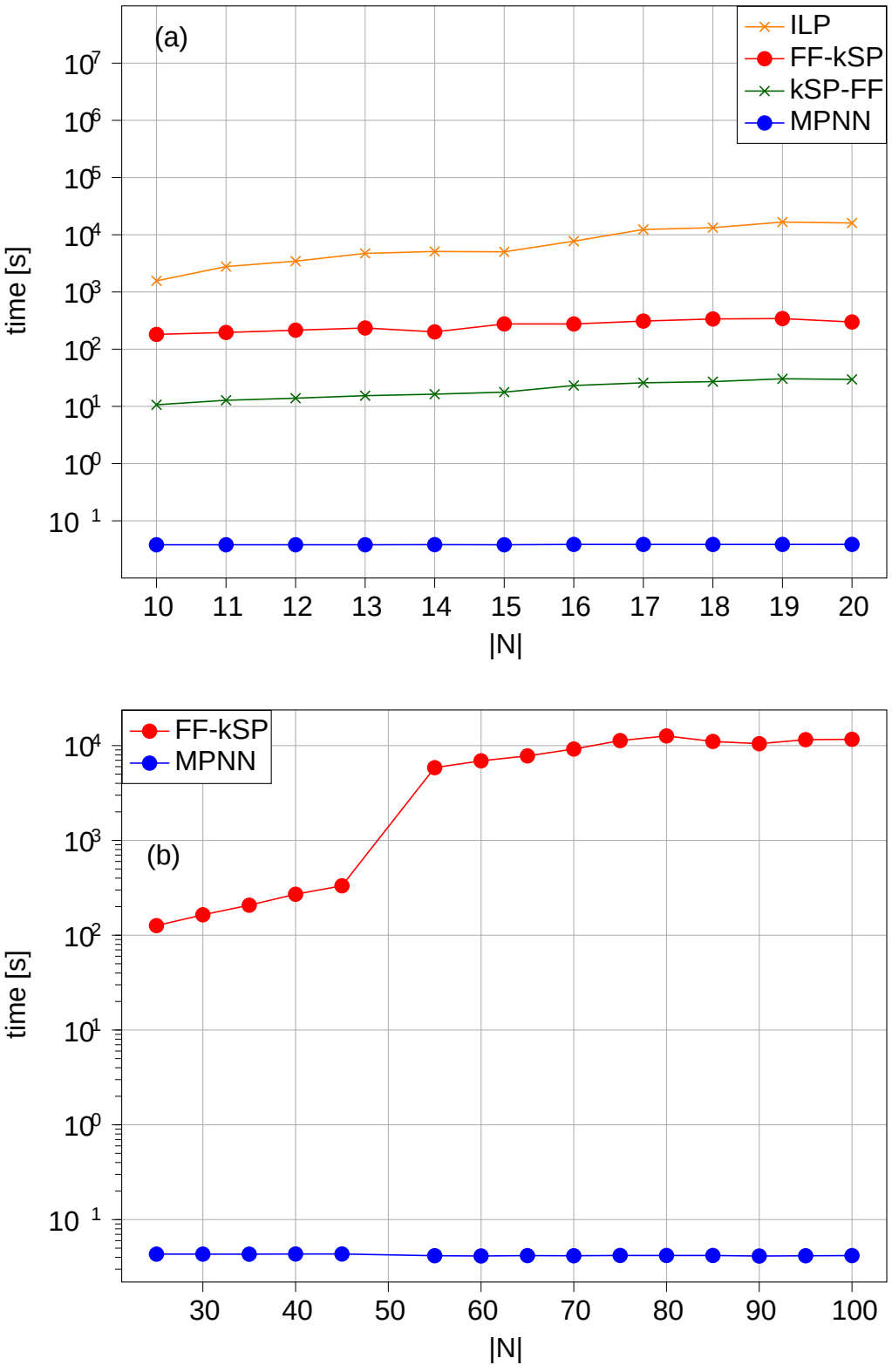


Fig. 4.11: (a) Comparison between computation times for throughput of networks over diffent node scales, using ILP, FF-kSP, kSP-FF and MPNN. (b) Computation times for throughput of networks using FF-kSP and MPNN [25] |N| = 100.

Graph Type	$_{TR}$	R^2		D_{KS}	F_{WSD}
ER	0.0	0.776	0.928	0.096	0.000286
BA	0.0	0.830	0.938	0.125	0.000201
SNR-BA	0.0	0.973	0.986	0.000	2.13 10 ^{6}
SNR-BA	0.2	0.901	0.983	0.079	2.13 10 ^{6}
SNR-BA	0.4	0.907	0.983	0.075	2.13 10 ^{6}
SNR-BA	0.6	0.906	0.983	0.060	2.13 10 ^{6}
SNR-BA	0.8	0.911	0.984	0.058	2.13 10 ^{6}
SNR-BA	1.0	0.907	0.982	0.062	2.13 10 ^{6}

Table 4.2: Accuracy for generalisation capability, measured by the coef cient of determination R^2 and . The variable $_{TR}$ determines how locally skewed the traf c is and D_{KS} is the absolute distance between the test throughput and the original training throughput (ks-2s test).

method for evaluating performance metrics of graphs, which are generally computationally dif cult to evaluate. This would allow for the fast evaluation of a large number of graphs within any topology design process. However, how well does this model generalise when varying the input structure and traf c distributions that affect the throughput of the network?

4.2.7 Generalisation Capability

The generalisation capability of an ML model refers to the ability of the model to operate over distributions not seen during the training process. In the context of this work, this could encompass different graph structures, different traf c distributions or distance scales. Here we choose to evaluate the model against two different graph structures and a change in input traf c distribution.

The generalisation capability is most probably the most important feature of this investigation, as without good generalisation, its dif cult to include this model in optimisation. As over the course of an optimisation procedure, one wants to evaluate networks with different traf c and structure, seeing which are the best performing structures and networks. If these are not accurate predicted values, then the optimisation is misguided.

ER and BA models - as described in section 3.2 - were used to test different graph structures[88, 84]. Using these models 5000 graphs with $25 \le |N| \le 45$ were generated per generative model. After calculating the maximum achievable throughput of these graphs using the FF-kSP heuristic and the Gaussian Noise (GN) model, we feed the graphs through the MPNN model to predict the throughput values.

When comparing the accuracy of the model over these varied graph structures, one can see that the accuracy, in terms of coef cient of determination (R^2), drops signi cantly, as seen in Table 4.2.

This means that the predictions vary largely from the actual labels of graphs. However, this is to be expected, as the graph structures resulting from the ER and BA models are largely different to those from SNR-BA graphs, as seen in section 3.5 [112]. This difference in graph structure can be quanti ed by the WSD [86], which measures the difference in structure between two sets of graphs and was de ned in section 3.1.3. When the WSD distance (E_{WSD}) is smaller, the graph structures are more similar. The WSD distance is calculated between the original SNR-BA test graphs and the various test sets and are shown in Table 4.2. Here one can see that the WSD is close to zero for the SNR-BA graphs, as they are generated from the same distribution as the originally tested SNR-BA graphs. We can see that the ER and BA graph's WSD distance is larger, where the ER structures are the most different from the original test graphs. In addition, due to the difference in graph structures, there are large differences in the throughput distributions. This is measured by the Kolmogorov-Smirnov two sample test, as used in sections 3.5 and 4.2.5, which returns a distance D_{KS}, which signi es the largest absolute difference between the CDF of two distributions. In Table 4.2, the large D_{KS} values for ER and BA test sets, signify the difference in throughput distributions to those seen during training. These D_{KS} values are larger than those of the SNR-BA test distributions, meaning their throughput distributions are different from those arising from SNR-BA distributions. The Pearson's correlation coef cient)(however remains high, which shows it still indicates the throughput trends of the networks well. This is important for optimisation, as the cost function might not need to accurately describe throughput exactly, however it needs to describe one graph being better than another.

We generated localised skewed traf c matrices for 5000 graphs to investigate how a change in the traf c demand matrix affects the accuracy of the model. By de ning the traf c as in Eq.(4.34), we generated 5 different traf c skews shown in Table 4.2.

$$\overline{T^{C}_{i;j}} = \frac{1}{\left(\sum_{k \in N} \frac{D(i;j)}{D(i;k)}\right)^{TR}} \tag{4.34}$$

For each of these skewed traf c matrices, we tested the MPNN accuracy in terms of R^2 and . The R^2 values drop by about 7%, and remain constant for the different skew values ($_{TR} > 0$). The Pearson's correlation coef cient)(value remain high at around 0.98, unchanged from the uniform traf c distribution results, meaning that it still predicts the throughput trend well.

The large variation in performance, in terms of R^2 for ER and BA graphs, indicates that the trained MPNN model does not generalise well to largely different graph structures. This highlights the importance of using a variety of different structures within the training dataset and that an expansion of the training data is necessary here

to represent different graph structures. Once the training set is more representative, by generating a variety of graph structures, it accurately evaluate the vast solution space of the topology design problem. This demonstrates the dif culty of training models that generalise to these vast solutions spaces. To achieve the modelling accuracy described in this section, a total of about 230,000 networks were used to train the MPNN. Given this huge data requirement of such graph representation learning frameworks, the natural question is whether there is some analytical graph representation, which can predict this topological performance well?

4.3 Analytical Network Throughput Representation

The previous section showed us that graph representation learning frameworks such as message passing neural networks are able to learn effectively from graph structured data, however capturing all the information that an optimisation process will encounter is inherently dif cult in the dataset generation process. This affects how accurately the model can predict on structures that it has never seen, as it simply does not know the relationship.

To simplify this process, one could simply omit the learning and derive a graph representation that incorporates features that affect the maximum achievable throughput of the network. Previously, research has focused on nding existing graph theoretic features that correlate to throughput or wavelength requirements [18, 26, 21], however deriving novel ones that correlate is not well researched yet.

An interesting starting point to this, is to look at work from sociology published by Leo Katz in 1953 [141]. Katz introduced the paradigm of analysing social networks in matrix or graph form. More speci cally, Katz was researching the idea of status of individuals and was focused on nding measures of centrality. He de ned a node-pair index value that aimed at understanding how well connected two nodes in a graph are connected as de ned in Eq.(4.35).

$$S_{Katz}[u;v] = \sum_{i=1}^{\infty} {}^{i}_{Katz} A^{i}[u;v] \tag{4.35}$$

Between each node pair $[u;v]$, one counts the number of paths available between them, weighting them by some decaying factor ${}^{i}_{Katz}$.

Another, independently developed idea of this measure is called the "communicability" [142]. This idea was developed for complex networks and followed the same idea as Katz, that shortest paths are not necessarily the only ones that in uence connectivity in the network. The communicability between nodes u and v is de ned by Eq.(4.36), where A is the adjacency matrix and i represents the path lengths that are being investigated.

$$S_{Comm}[u;v] = \sum_{i=1}^{\infty} \frac{A^i[u;v]}{i!} = (e^A)_{uv} \tag{4.36}$$

Here the communicability also weights longer paths less. Strictly both of these measures use walks and not paths, meaning that the nodes (except for source and destination nodes) can be repeated. Using these metrics as a starting point, the question of how one can incorporate the topologically important features and inputs of optical networks and their maximum achievable throughput in a graph level metric, was investigated.

4.3.1 Demand Weighted Cost

For optical networks, it is well-known that path diversity is important within routing performance, however also resilience within the network. In addition, the quality of a path affects the performance of a network, i.e. how long that path is, in terms of its hops and its physical length. The topology of the network determines these path properties, however it is also important that these paths exist relative to the traf c between the source destination nodes, with a lot of short diverse paths between high traf c source destination nodes being a bene cial characteristic. These characteristics are termed communication costs and are de ned in Eq.(4.37) and Eq.(4.38). Here $H_k^{DWC}[u;v]$ is the number of hops for path k between nodes u and v and $L_k^{DWC}[u;v]$ is the length of path k between nodes u and v. $_k^{DWC}$ represents a weighting for how important each kth-path is within the metric. By altering the topology of the network, one can change the characteristics of these paths and therefore also the routing solution of the network. This directly impacts the maximum achievable throughput within networks.

$$C_H^{DWC}[u;v] = \overline{T_{uv}^C} \sum_{k \in K_{uv}} {}_k^{DWC} H_k^{DWC}[u;v] \tag{4.37}$$

$$C_L^{DWC}[u;v] = \overline{T_{uv}^C} \sum_{k \in K_{uv}} {}_k^{DWC} L_k^{DWC}[u;v] \tag{4.38}$$

By weighting the sum by the relative traf c between that node pair and counting the path length in terms of hops and spans, it is possible to incorporate the major topological features and inputs of optical networks in a centrality measure. How can we aggregate these node pair features and summarise these on a graph level?

Simply taking the sum over all node-pairs, i.e. a common aggregation strategy, as in neural message passing, gives a graph-invariant aggregation. Terming the resulting graph representation as the demand weighted cost (DWC):

$$C_{DWC} = \sum_{u,v \in Z} C_L^{DWC}[u;v] + (1 \quad) \; C_H^{DWC}[u;v] \tag{4.39}$$

where is a constant that de nes the importance of structural and physical properties of the paths respectively. By weighting the path costs by the constant, the degree to which the number of hops or physical length of paths impact the metric is determined. Chapter 3 showed that both the structural and physical properties impact the maximum achievable throughput. However, the impact relative to each other was not discovered.

Path lengths, in terms of physical distance, directly impact the maximum achievable throughput of optical networks as the achievable throughput of any lightpath is equivalent to $2B_{CH}\log_2(1 + SNR_z)$, where SNR_z is the SNR for a lightpath between z. The SNR is majorly reliant on the number of spans that a lightpath traverses. Although the SNR also relies on the congestion along the path, the congestion is often close to fully occupied when the network is operated at its maximum load. On the other hand, path lengths - in terms of hops - do not have a direct impact on the maximum achievable throughput. As was seen in Chapter 3, the SNR-BA graphs with longer diameters had larger wavelength requirements, compared to the ER/BA graphs with shorter diameters and lower wavelength requirements. Generally, low diameters (smaller path lengths in terms of hops) are desirable, as this was shown to give lower wavelength requirements. This, however, needs to be balanced with short physical path lengths to maximise achievable throughput. Both physical path lengths and path lengths in terms of hops do not fully determine the maximum achievable throughput and therefore the DWC is a graph metric aimed at capturing most, however not all the factors that determine the maximum achievable throughput. Using these path metrics however is bene cial, as the computational complexity relies on only the scaling of path computation, which scales linearly.

The communication costs in terms of hops (C_H^{DWC}) and bre spans (C_L^{DWC}) were visualised, by using a 15 node network with node locations sampled uniformly from a grid and generated via the SNR-BA model is investigated. Uniform traf c between all node-pairs was used to calculate both C_H^{DWC} and C_L^{DWC} and plotted in Figure 4.12. The cost was calculated by using $k = 3$, constraining $_k^{DWC} = 18k \in K_{uv}$, meaning that all k-paths are weighed equally in the cost. In both gures 4.12a and b the lighter patches refer to larger values of the cost, meaning that there is a higher product between the traf c and path lengths. These lighter patches in the network are parts of the network that are at a path disadvantage compared to the darker patches in the network that are better connected relative to the traf c that needs to ow between those nodes. Empirically both C_H^{DWC} and C_L^{DWC} are similar in the patterns that they show, however there are subtle differences.

For example, looking at the top left hand corner, for (a) the hotspot is between node-pair (15,1) and (14,1), for (b) these are also high valued areas, meaning that the paths are both longer in hops and physical distance. However, for node-pair (5,13) in Figure(a), exhibits a hotspot, whereas in Figure(b), the same area is in a darker area,

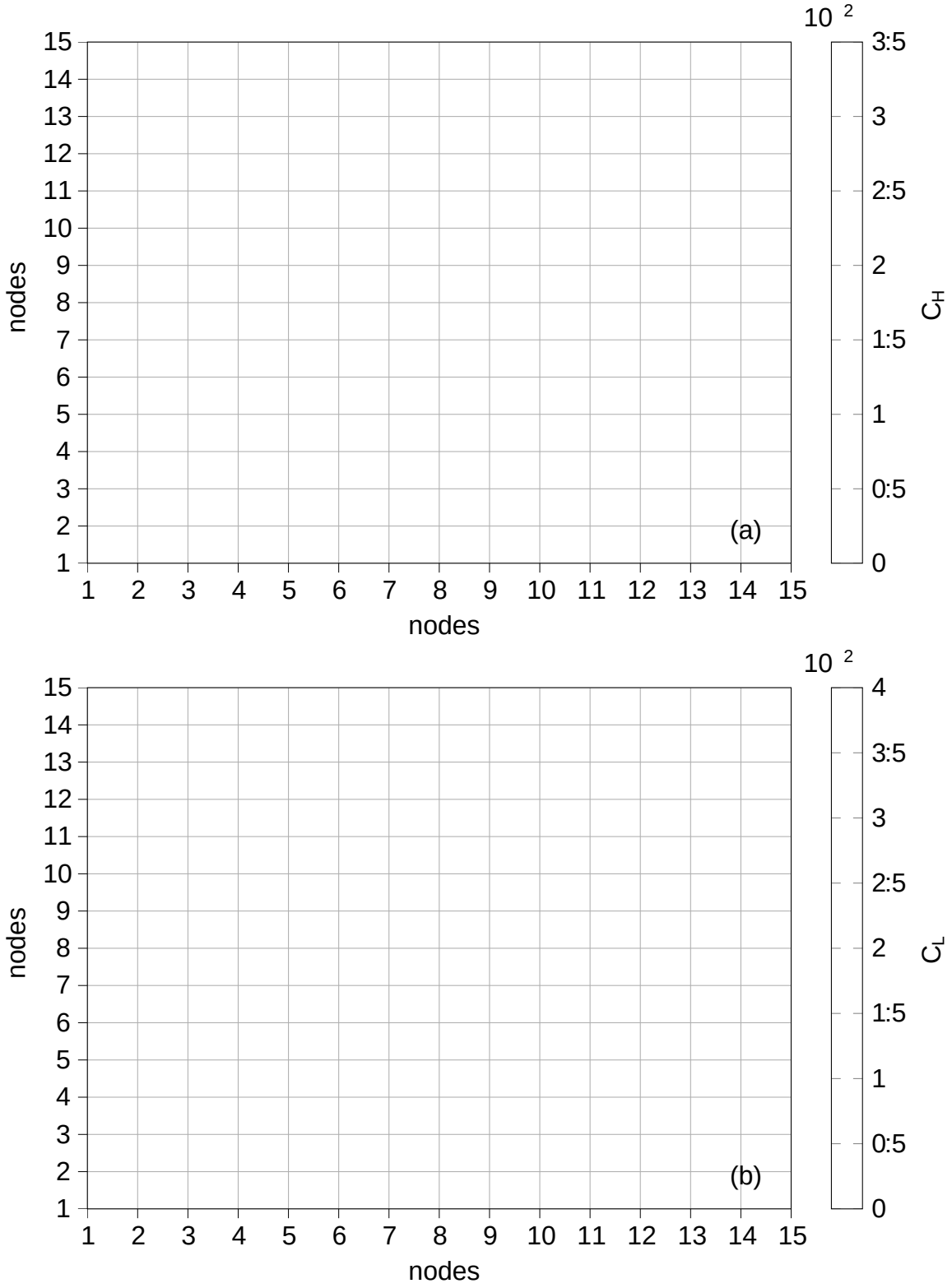


Fig. 4.12: Contour plot showing communication cost in terms of (a) hops (C_H^{DWC}) (b) fibre spans (C_L^{DWC}) between node pairs in a 15 node optical network generated by the SNR-BA model and uniform traffic.

	0.64	0.64	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.86	0.86
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4.3: Pearson correlation coefficient for C_{DWC} for different α values correlated against ILP throughput of 6000 SNR-BA graphs with $|N| = [10, 15]$.

meaning that the paths between (5,13) are among the longest paths in terms of hops, however, not in terms of physical length. This contrast in terms of structural/physical cost in the network begs the question: which cost is more important in terms of the maximum achievable throughput? This is a similar problem that was briefly discussed in Chapter 3, that ER/BA graphs had better structural properties (smaller wavelength requirements and smaller diameter), however the physical path lengths were very large.

Measuring the contribution of both the structural and the physical contribution to maximum achievable throughput is difficult in reality. To understand what value to choose for α in Eq.(4.39), the graphs used in section 4.2.3 have been re-used. 6,000 graphs from the alpha dataset with nodes $10 \leq |N| \leq 15$ were investigated. The C_{DWC} values were calculated with $\alpha = [0, 1.0]$ and these are then correlated against the maximum achievable throughput values calculated via the ILP (Pearson's correlation coefficient). These values are then summarised in Table 4.3. It turns out that many values have very similar correlations, this being simply down to the large similarity between most paths in terms of their hops and physical path length. However, when $\alpha < 0.2$ one can observe that the correlation drops sharply. This indicates that when heavily weighting only the path length by hops it does not correlate well to the actual maximum achievable throughput. From the results in Table 4.3 and Chapter 3, it is clear that physical path lengths play a role within maximum achievable throughput.

Observing that as $\alpha \geq 0.2$ the correlation stays largely the same, a value of $\alpha = 0.5$ was chosen to test the metric on the other two test datasets generated in section 4.2.3. Here 15,000 graphs were tested with node scales $25 \leq |N| \leq 100$ evaluated for their C_{DWC} using $\alpha = 0.5$. These results are then plotted as scatter plots in Figure 4.13, with their corresponding Pearson's correlation coefficients (r). Here one can clearly see that C_{DWC}^{1} has a strong linear correlation to the maximum achievable throughput of that of FF-kSP, even more so than to that of the ILP.

One of the problems that was observed with learning graph representations for maximum achievable throughput was that the MPNN models struggled to generalise to different graph structures, which is vital within the optimisation process. The same ER and BA graphs are tested to investigate whether DWC works well on different structures too. Calculating DWC and correlating this to the maximum achievable throughput of each of these datasets shows a strong correlation in both ER and BA sets of 0.98. This demonstrates that DWC correlates strongly within different graph structures too.

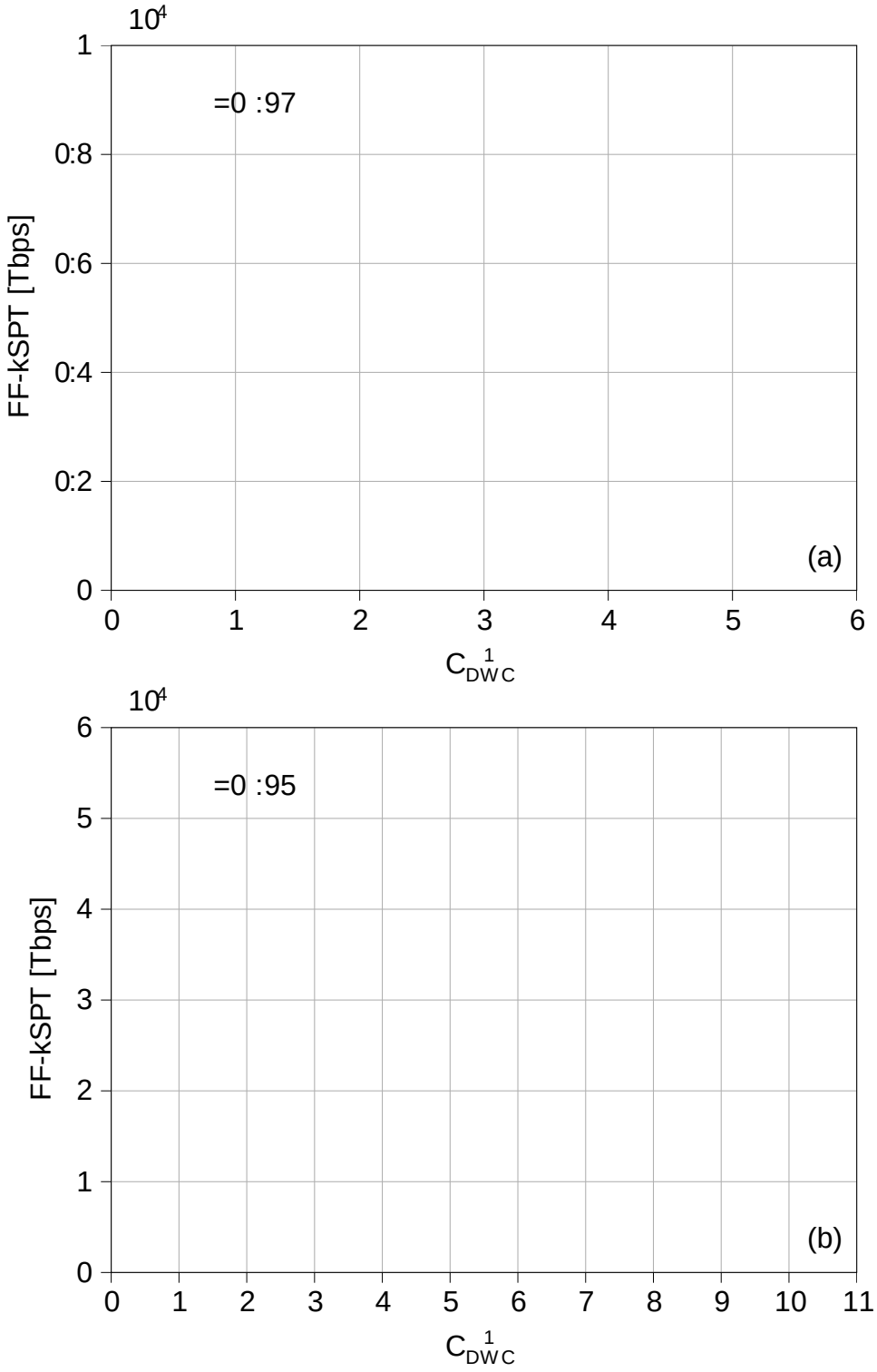


Fig. 4.13: (a) DWC correlation, calculated with = 0 :5, to the FF-kSP maximum achievable calculation for 25 j N j 45 (5,000 samples). (b) DWC correlation, calculated with = 0 :5, to the FF-kSP maximum achievable calculation for 55 j N j 100 (10,000 samples).

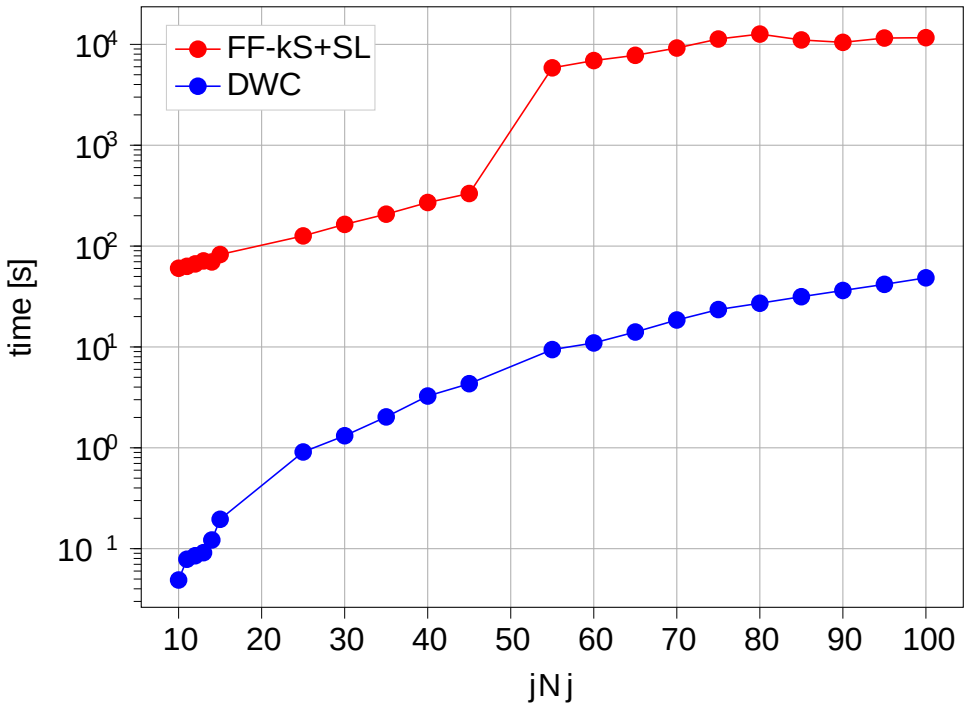


Fig. 4.14: Computation time for computing C_{DWC} compared to using FF-kSP + Sequential Loading (SL) heuristic to assess topology throughput performance.

The time required to calculate the performance of graphs via FF-kSP+SL and DWC was measured to demonstrate the complexity scaling and plotted in Figure 4.14. It can be clearly seen that DWC scales approximately the same way as the heuristic, however with a nearly constant lower time shift of approx two orders of magnitude. This means that this method can be applied within optimisation methods during the design process of the network and enables a computationally ef cient graph representation that correlates well with maximum achievable throughput of optical core networks.

4.4 Summary

This chapter investigated different methods of approximating and calculating the upper bounds of the maximum achievable throughput and nding/learning graph representations that can correlate to this performance metric, all with the objective of including it in future optical core network optimisation. Initially, upper bounds on the maximum achievable throughput were derived, however they depend on the optimal path (and wavelength) distributions. It was shown that actually constraining the resources at edge level and relaxing the integer wavelength constraint can give a very good approximation of the maximum achievable throughput when neglecting physical

properties. This was previously stated by David Ives [122], however was not shown how large a deviation this is from the actual optimal values. The problem of needing to solve a linear programming problem remained. Although computationally efficient, it is not efficient enough to include in the combinatorial optimisation of optical networks, which requires targets that can be determined in ms time.

With the objective of reducing the computational complexity, the idea of learning graph representations has previously been exploited in other communications and quantum chemistry research especially. Motivated by this work an investigation of whether message passing neural networks can learn these graph representations to correlate to maximum achievable throughput of networks of varying size. It was shown that message passing neural networks have this capacity to learn over varying network sizes, given 100s of thousands of training network samples. However, generalising to different structures that were unseen during training is difficult. The problem, therefore, remains that given that one is trying to find possibly new structures in the networks that have not been seen before, these might be falsely labelled in the optimisation process. Therefore, it seems not as straightforward to learn completely generalisable models of maximum achievable throughput, without labelling huge amounts of data that represent the full variation in structures that one wants to optimise over.

Previously, most researchers looked at using metrics from complex network research, such as algebraic connectivity and number of spanning trees to correlate these measures to wavelength requirements in optical networks. Inspired by this, an investigation into whether one can derive a graph representation from previous complex network research that has focused on network problems that depends on communication of node-pairs in a network. By investigating two measures that have previously investigated such communication, i.e. the Katz-index and communicability, an optical network based graph metric was defined, denoted as the demand weighted cost. Comparing this metric to the datasets used to evaluate the MPNN, it was shown that it has, especially for FF-kSP, a strong correlation for the maximum achievable throughput with two orders of magnitude less computational cost.

Therefore, demand weighted cost reduces the computational complexity by 2 orders of magnitude, whilst retaining a linear correlation to maximum achievable throughput with a Pearson's correlation of $= 0:98$. A strong linear correlation, however does not guarantee a causitive effect. Therefore, the next step was to investigate whether intentional reduction in demand weighted cost maximises achievable throughput within physical topology designs of optical networks. This is investigated in the following and final work Chapter 5.

Chapter 5

Maximising Throughput in Physical Network Design

Prior to the start of this PhD research, the physical topology design (PTD) problem focused on wavelength requirements, resilience, latency and cost. Maximum achievable throughput was initially only investigated from an operational point of view, however not included in the PTD problem [75, 76]. As seen in Chapters 3 and 4, both the impact of structural and physical properties, as well as methods for including the maximum achievable throughput in the PTD problem were investigated. In addition to this, other works have made similar investigations [20, 21, 7, 135]. However, to-date there has not been a set of research that has directly optimised the maximum achievable throughput of optical networks.

Chapter 4 concluded with DWC being highly correlated with the maximum achievable throughput of an optical network in various settings. This chapter investigates if this correlation can be used in PTD of optical networks to maximise the maximum achievable throughput. In this Chapter, the focus is on the combinatorial optimisation of the physical topology, and therefore the methodologies used to solve the PTD problem are initially investigated.

5.1 Integer Linear Programming Formulations for Physical Topology Design

ILP formulations are one of the only methodologies to achieve provably global optimal topology design. Both in Chapters 3 and 4 ILPs were used to calculate the maximum achievable throughput of a network, by optimising the RWA given a traf c distribution. In this section, ILPs are discussed in the context of combinatorial optimisation for the PTD problem.

For the PTD of optical networks, there have been a couple of ILP formulations developed [11, 10]. The rst of which was a ow formulation, meaning that it was formulated according to the ows of demands across the network, by relaxing the wavelength continuity constraint [11]. The adjacency matrix entries $a_{uv}(i,j)$ is de ned as a variable as in Eq.(5.1) and whether a demand between nodes i and j traverses over edge$(u;v)$ is de ned by the variable$^{ij}_{uv}$ in Eq.(5.2).

$$a_{uv} = \begin{cases} 1 & \text{if nodes } u \text{ and } v \text{ are connected via a fibre} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

$$\lambda_{uv}^{ij} = \begin{cases} 1 & \text{if demand between } i \text{ and } j \text{ uses edge } (u,v) \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

The conservation of flow is ensured by Eq.(5.3), where D_{ij}^{LP} is the number of requested lightpaths between nodes i and j.

$$\sum_{v \in N} \lambda_{uv}^{ij} - \sum_{v \in N} \lambda_{vu}^{ij} = \begin{cases} D_{ij}^{LP} & \text{if } u = i \\ -D_{ij}^{LP} & \text{if } u = j \\ 0 & \text{otherwise} \end{cases} \quad \forall u \in N, (i,j) \in Z \tag{5.3}$$

As each fibre in an optical network has a limited number of wavelengths for transmission, the number of wavelengths used on each edge is constrained by Eq.(5.4), which constrains the number of flows over each edge to the value of W_j, defined as the number of wavelengths available per fibre. In this formulation W_j is kept as a variable to be optimised and not a constant.

$$\sum_{i,j \in Z} (\lambda_{uv}^{ij} + \lambda_{vu}^{ij}) \leq W_j \quad (u,v) \in Z \tag{5.4}$$

Lightpaths only should be routed over existing edges and therefore are constrained by Eq.(5.5), where G is a large constant. This constraint ensures that feasible routings are achieved by the designed network.

$$\sum_{i,j \in Z} (\lambda_{uv}^{ij} + \lambda_{vu}^{ij}) \leq G \, a_{uv} \quad (u,v) \in Z \tag{5.5}$$

Finally the topology is constrained to have a total fibre distance of L_{NET}, ensured by constraint defined in Eq.(5.6), where $D_{u,v}^{fibre}$ is defined as the distance between nodes u and v.

$$\sum_{u,v} a_{uv} \, D_{u,v}^{fibre} \leq L_{NET} \tag{5.6}$$

The overall objective in this ILP is to minimise the number of wavelengths, defined in Eq.(5.7). This is the objective that was studied intensely in the late 1990s and early 2000s [18, 93, 26]. This formulation is still relevant however, as it could be extended to maximising the number of lightpaths too, by integrating the ILP described in section 4.1.1.

$$\min(|W|) \tag{5.7}$$

The second ILP, is a simplified version of the flow formulation [10]. It aggregates flows at the source nodes, therefore termed the source-formulation. This aggregation simplifies the ILP problem, as instead of having to iterate over $\frac{|N|(|N|-1)}{2}$, the source-formulation iterates over $|N|$ source nodes. The flow formulation has a computational complexity that grows with $O(|N|(|N|-1)|E|)$, whereas the source formulation grows with $O(|N||E|)$.

$$^{i}_{uv} = \begin{cases} 1 & \text{if demands with source } i \text{ use edge } (u,v) \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

The source formulation defines whether a lightpath starting at node i traverses edge (u,v) with $^{i}_{uv}$, defined in Eq.(5.8). Therefore, instead of iterating over each node-pair, the source formulation iterates over each source to ensure the flow of lightpaths, as defined in Eq.(5.9).

$$\sum_{u \in N} \sum_{v \in N, v \neq u} {}^{i}_{uv}\, a_{uv} = \sum_{j} D^{LP}_{ij} \quad \forall i \in N \tag{5.9}$$

The constraint defined in Eq.(5.10) ensures the routing of all lightpaths, with Eq.(5.11) constraining the number of wavelengths used. The previous distance constraint (Eq.(5.6)) is used to constrain the fibre distance. The same objective of minimising wavelengths can be applied here too (Eq.(5.7)).

$$\sum_{u \in N} \sum_{v \in N, v \neq u} {}^{i}_{uv}\, a_{uv} = \sum_{u \in N} \sum_{v \in N, v \neq u} {}^{i}_{vu}\, a_{vu} \quad D^{LP}_{ij} \quad a_{vu} \quad \forall (i,j) \in Z \tag{5.10}$$

$$\sum_{i \in N} {}^{i}_{uv}\, a_{uv} \leq |W| \quad \forall (u,v) \in Z \tag{5.11}$$

These ILP formulations were restricted to small networks (10-15 nodes) with few wavelengths (8).

As seen in Chapter 4, the ILP formulation is not feasible for networks larger than 20 nodes for routing, let alone network design. Therefore, to-date it is not a feasible solution to the PTD problem for modern networks. To find solutions to the PTD problem, other heuristic ways of traversing the solution space were investigated in the course of the research work described in this thesis.

5.2 Heuristics/Meta-Heuristics for Physical Topology Design

Due to the computational complexity of designing WRONs, many heuristics and meta-heuristics have been applied to the problem. The simplest of which would be greedy heuristics, which apply sequentially the best decision for each edge addition, however do not look at the best combination of edges to add [143]. These solutions generally get stuck on local optimas as they do not explore the solution space effectively [143].

As within the design for distributed computer networks, branch exchange is an algorithm that has been applied to the PTD problem of WRONs [144, 145]. Here an edge (or edges) is added to the network given some objective, whilst also removing an existing edge (or edges) from the network. After this pertubation the performance of the network is quanti ed and if the performance is better, the change is accepted. Once all the possible topology transformations are explored, the search process is stopped. This procedure however also relies on local transformations to try and achieve global optimas, only possible in designing tree networks [73].

Cut saturation is another heuristic which has been applied to both distributed computer networks design as well as within WRONs [146]. This method can be considered as an extension of the branch exchange heuristic, however instead of looking at all possible topology transformations, it looks for the transformations that are likely to improve on the objective function. This is achieved by calculating the network cut, that severs the network into two sub-networks over the fewest, most congested edges. Once this is found, another edge is added across the cut boundary. This however is dif cult in practice, since evaluating network cuts is computationally expensive for large networks and evaluating the saturated cut requires a knowledge of the routing, an NP-hard problem within optical networking [35]. Although more recently this has been achieved for moderate sized networks ($N = 20$) [19] and is a promising new research area.

Previously, generative graph models have been used as heuristics for designing WRONs [147]. In speci c, an altered version of the Gabriel graph model has been used to compare against a multi-objective evolutionary algorithm (MOEA). In Araujo et al [147], the generative graph model offered some solutions that were better than the MOEA for larger networks (68 nodes) at 1% of the computational run time. For small networks (18 nodes) the MOEA still outperformed the generative graph model with 55% lower blocking at the same cost. Generative graph models generate networks with speci c structures and physical properties and do not optimise a speci c optimisation goal. Shown in Chapter 3, the SNR-BA graphs tended to produce well-performing physical properties and therefore outperformed other generative graph models. Which optimisation methods can maximise achievable throughput well, is one of the goals of this chapter.

In the 2000s meta-heuristics were - in addition to hand-crafted heuristics and ILP formulations - researched for the design of WRONs. Evolutionary algorithms (EA)/genetic algorithms (GA) and particle swarm optimisation (PSO) being the most applied meta-heuristics to date [148, 149, 150, 151, 152, 153, 154]. Greedy/Cut-saturation heuristics are single solution heuristics, meaning they improve on a single solution over the course of optimisation. This means they are particularly vulnerable to getting stuck in local optima. Both GA and PSO methodologies work with populations of solutions, where each individual represents a binary encoding of a solution. This approach helps with diversi cation of the search space, as not only a single solution is explored, however a population of solutions.

PSO is based on the observational theory of bird ocking, sh schooling and swarming theory and was initially proposed by Eberhart and Kennedy in 1995 [155]. The main idea is that a population of particles (solutions) are randomly initialised with a random velocity associated with them. Each particle traverses the search space and records its personal best and in total the global best too. It then updates the velocities according to these best values to converge globally on a best value, whilst exploring different regions of the search space.

GAs operate on the evolutionary principle by encoding binary solutions in the population as genetic vectors [156]. Parents are selected from the population to reproduce based on an objective function. The process of swapping parts of these solutions during the reproduction is termed crossover, after which random mutation occurs with a pre-set constant probability. The offspring then replace the worst individuals in the population and the process continues in search of better solutions with each iteration.

Naturally GAs suit the nature of binary optimisation problems better than PSO and, have, therefore, been explored in much greater depth for the PTD problem [150]. This is because PSO naturally explores continuous solution spaces and can be adapted to discrete optimisation [157]. In [158] both GA and PSO were applied to the logical design of optical networks, with the GA design allocating approximately 50% more traf c at a similar cost to the PSO-designed network. Given that the GA algorithm is originally an encoding of a discrete optimisation problem, it is selected as the meta-heuristic of choice for this Chapter.

The ILP, meta-heuristic and heuristic methods all attempt to solve the combinatorial optimisation problem behind designing optical networks. However, none of these methods have directly optimised the maximum achievable throughput. To investigate whether the previously suggested intermediary objective function of DWC can maximise the achievable throughput of optical networks, the next section incorporates it within a generative graph model, greedy and genetic algorithm design methods and the resulting maximum achievable throughput is investigated.

5.3 Maximising Achievable Throughput in Physical Topology Design

Designing optical network to maximise achievable throughput, requires the simultaneous solving of both NP-hard combinatorial optimisation problems of PTD design and RWA problem. To combat this computationally dif cult problem, graph properties that have been seen to correlate highly to maximum achievable throughput are summarised via a metric termed DWC. DWC must be included in the design of optical networks and compared against a controlled dataset to prove whether it can result in the maximisation of achievable throughput. This section investigates the DWC metric as a proxy for maximising achievable throughput in optical networks. This is done by using generative graph models, greedy and genetic algorithms to optimise networks in terms of their DWC. Then evaluating these designed networks via an ILP formulation to calculate their true maximum achievable throughput. The next section details the methodologies used to design topologies and investigates the different design methods.

5.3.1 Dataset Generation

Whether optimising DWC leads to higher throughput networks was tested by generating a set of nodes to perform PTD optimisations using DWC as the objective function. These nodes are generated by scattering nodes over a rectangle the size of the north-American continent with equal probability, as in section 4.2.3. 10 sets of 20 nodes each are generated as PTD problem instances. These sets of nodes are then used as inputs to the design algorithms described in the following sections. For each of the algorithms the goal was to achieve a connectivity of approximately $\bar{\nu} = 0.15$ (de ned in section 2.3 Eq.(2.19)), which equates to about 28 edges for a 20 node network.

In section 3.6.2, the throughput of SNR-BA networks was seen to perform well compared to ER, BA and real networks. Therefore, the SNR-BA model introduced in Chapter 3, is used to generate a set of networks independent of the DWC metric. These are used as a control-set, to determine whether there is a signi cant shift in achievable throughput of the designed optical networks. The SNR-BA-random networks are referred to as the control-set within this chapter.

With that goal in mind 10 SNR-BA networks were generated for each of the generated node-sets and are termed SNR-BA-random.

In addition to this control-set, an additional set of SNR-BA networks were generated. These networks were generated by generating 100 networks for each set of nodes. The DWC was evaluated for each of these 100 networks and the network with lowest DWC was chosen. This is done to minimise the DWC of these networks. This set of networks is termed SNR-BA-DWC, as they were chosen according to their DWC values.

As introduced in section 5.2, greedy heuristics are the simplest of heuristic optimisation algorithms. Within the greedy heuristic, the PTD problem is modelled as a sequential decision making problem. At each timestep the DWC of each possible edge choice is evaluated. The edge that achieves the lowest DWC at each timestep is chosen. This procedure is followed until the desired edge number is achieved. To ensure that bi-connectivity is achieved, a simple ring network is constructed from the node-set before beginning the greedy heuristic, which is then given as a base-topology. This means that a part of the edge choice are also independent of DWC, however this is necessary to ensure the bi-connectivity constraint of optical networks. These networks are referred to as the Greedy-DWC networks.

As introduced in section 5.2, GAs de ne a set of solutions as a population. This consists of a set of *genomes*, which are binary vectors encoding solutions to the optimisation problem. In the case of PTD, this binary vector is de ned as the upper triangle of the adjacency matrix. An objective function value is associated with each of these individuals in the population. The best performing individuals are used to parent new solutions for the future population. The parents swap a portion of their genome (crossover) and mutate their genomes according to pre-de ned probabilities. The objective values of these new genomes are then calculated and replace the worst performing individuals in the population, if found to be better. This procedure is illustrated in Figure 5.1. After many iterations, the idea is that the objective function is optimised as the population performs better.

The GA was evaluated by generating an initial population of 10,000 networks by creating a ring network (to ensure bi-connectivity) - as was done for the greedy heuristics - and then adding random edges until a connectivity of $0.15 \times (28)$ is achieved. The GA was run for 500 iterations, higher iteration number is preferable, however not possible due to time constraints. The mutation probability of 0.1 was chosen to ensure diversity of the new solutions and a crossover probability of 0.8. At each iteration 30% of the population is used to parent children. These values were seen to perform well from previous research in [159]. The resultant topologies are termed GA-DWC.

The search space of this problem is extremely large. To evaluate edge choices within a 20 node topology, there are 190 edges possible edges ($\frac{N(N-1)}{2}$), 170 if a ring topology already exists, for which there are 1.5×10^{12} possible combinations to choose 8 from the 170 edges. Therefore, in addition to the above method, where the full upper triangle of the adjacency matrix is used to optimise the topology, a smaller optimisation problem is de ned. Here the initial population is generated by only creating ring topologies again, however without adding random edges. The binary encoded vectors in this case are not the upper triangular vector, however a vector of length 100. Each entry corresponds to a pre-de ned edge. These 100 edges were chosen by the same greedy algorithm as

Fig. 5.1: Process of the genetic algorithm: parents are selected from the population, after which crossover and mutation are performed and the new solutions (children) are added back into the population upon evaluation.

described in section 5.2. The other optimisation parameters were kept the same. These networks are termed GA-DWC-Greedy, as a greedy algorithm was used to find the edge choices.

5.3.1.1 Physical Topology Design Objective Evaluation

Due to its computational simplicity, DWC was used as a intermediary objective function for the PTD problem. The DWC consists of both structural and physical path length costs. Chapters 3 and 4 showed that both the structural and physical properties impact the maximum achievable throughput of optical networks. However, the degree to which structural or physical properties impact the maximum achievable throughput is difficult to gauge. On one hand, a network requires a small diameter in terms of hops. This means that lightpaths will use less wavelength resources within the RWA, as fewer edges need to be traversed. However, if this small diameter comes at a cost of large physical path lengths, the gain made by saving wavelength resources is lost by a poor achievable SNR of the additional lightpaths. This was seen for example in Chapter 3, where the ER and BA networks showed increased number of lightpaths assigned, however with reduced throughput per lightpath. Understanding which of these is a priority within the optimisation is important for maximising the achievable throughput in optical networks.

By making the assumption that the NLI in optical networks accumulates

incoherently, the SNR scales approximately linearly with the inverse of the number of spans (n_{SP}). However, as the SNR is included in the logarithm of throughput (Shannon capacity), the throughput of these lightpaths reduces with the logarithm of the number of spans. However, maximum achievable throughput is the sum of throughput of all lightpaths routed, therefore by increasing the number of lightpaths routed occurs outside of the logarithm of Shannon's capacity.

Therefore, one can clearly see that there is a trade-off to be made between the relationship of structural and physical properties of a network. Which is more important? Increasing the number of lightpaths allocated, or the resulting physical properties of these lightpaths? Does allocating more lightpaths lead to larger physical path lengths? If so, at what rate?

The weighting of the DWC was evaluated at values of 0 and 0.5 to understand this trade-off better. A value of 0 represents an optimisation to minimise the path length purely in terms of hops. A value of 0.5 represents an equal weighting between the structural path lengths and physical path lengths. The value of $= 1.0$ was omitted as it did not show any new insights compared to $= 0.0$ and $= 1.0$.

5.4 Designing Topologies using Demand Weighted Cost

5.4.1 Demand Weighted Cost Minimisation using $= 0$

The exact trade-off between structural and physical properties was initially investigated using the set of nodes generated with a design objective of minimising DWC with $= 0$. Ten topologies per optimisation algorithm: SNR-BA-DWC, Greedy-DWC, GA-DWC-Greedy and GA-DWC methods, were designed. These were then compared to the control-set networks, generated via the SNR-BA-random method, which does not use DWC to generate the network topology. In total, this resulted in the comparison of 50 topologies. After generating these topologies their maximum achievable throughput was calculated using the ILP and PLI models, defined in sections 4.1.1 and 3.6.2, respectively. These maximum achievable throughput values were then plotted in Figure 5.2. The plots in Figure 5.2, are violin plots, with traditional box-plots in the middle, including median, interquartile range and the outer-most lines signifying 1.5 times the interquartile range of the data. The density estimation of the distribution of maximum achievable values is drawn on top of this box-plot in colour.

It is clear from Figure 5.2 that all design methods have a significantly increased average maximum achievable throughput, with the largest being the greedy algorithm delivering an increase of 81.5% over the control-set. The average DWC, number of lightpaths allocated and maximum achievable throughput of networks were calculated and shown in Table 5.1. The DWC of all the design methods is clearly minimised

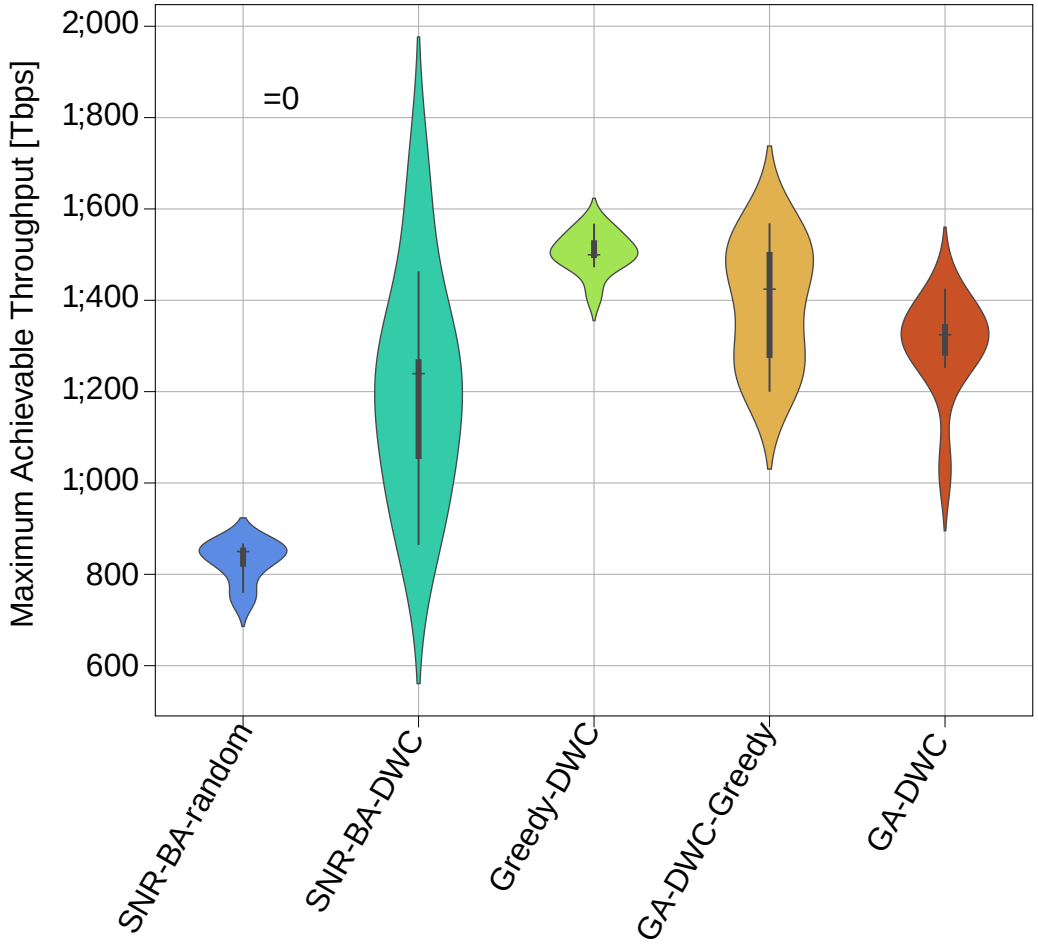


Fig. 5.2: Resultant maximum achievable throughput of networks designed using SNR-BA, Greedy and GAs to minimise the DWC at $= 0$, compared to the control-set of SNR-BA-random.

compared to that of the control-set (SNR-BA-random). However, when comparing the average DWC values amongst the design methods, it does not necessarily give a clear performance trend. The best performing networks are those designed via the greedy-DWC method. However, the smallest DWC values on average follow from the GA-DWC methods. To understand better, why the greedy-DWC methods perform better than the GA-DWC methods, the impact of both the structural versus the physical properties of the networks was investigated further.

The maximum achievable throughput depends on two properties of the network. Firstly, the number of allocated lightpaths depends on the limiting cut in the network, de ned in Chapter 4 in Eq.(4.2). Given a certain traf c distribution, the limiting cut within an optical network dictates the maximum number of allocatable lightpaths. Secondly, the physical properties of the path used by these allocated lightpaths determines the achievable throughput of these lightpaths. Therefore, there is a balance

Method	DWC	LP allocated	Throughput [Tbps]	R_T	R_P	R_{LP}	
SNR-BA-random	4.98	1482	828.38	1.0	1.0	1.0	0.0
SNR-BA-DWC	0.39	2185	1208.26	0.98	1.02	1.47	0.0
Greedy-DWC	0.35	2850	1504.17	0.92	1.25	1.92	0.0
GA-DWC-Greedy	0.37	2622	1393.33	0.93	1.23	1.76	0.0
GA-DWC	0.33	2793	1298.94	0.83	1.70	1.88	0.0
SNR-BA-random	7.30	1482	828.38	1.0	1.0	1.0	0.5
SNR-BA-DWC	2.85	1691	923.16	0.99	0.98	1.14	0.5
Greedy-DWC	2.74	2356	1295.90	1.00	0.94	1.58	0.5
GA-DWC-Greedy	2.86	2166	1178.93	0.98	0.99	1.46	0.5
GA-DWC	2.98	2603	1366.85	0.99	1.04	1.75	0.5
Greedy-Cut	0.40	3325	1711.32	0.98	1.07	2.14	0.0
RL-Cut	0.39	3971	2055.51	0.92	1.29	2.68	0.0

Table 5.1: Average values for DWC, lightpaths (LP) allocated, maximum achievable throughput, worst-case achievable throughput ratio (R_T), path length ratio (R_P), lightpaths allocated ratio (R_{LP}) and values for all designed topologies.

between increasing the number of lightpaths that can be allocated, whilst not increasing their physical length too much. Three ratios were proposed as part of the research in this thesis to understand the trade-off between network structure and physical properties. These ratios were then investigated for the topologies of each design method. These ratios attempt at splitting the achieved throughput of networks into the allocated lightpaths and the path lengths and investigates the effect of both of these on the resultant maximum achievable throughput.

Initially the average SNR between each source and destination was calculated using a fully loaded centre channel on each of the traversed edges. The resultant achievable throughput over this lightpath is calculated as T_z, defined in Eq.(5.12).

$$T_z = 2 \quad B_{CH} \quad \log_2(1 + \overline{SNR}_z) \tag{5.12}$$

This value is then summed over all source-destination pairs and expressed as a ratio in Eq.(5.13). This is the ratio of average worst-case achievable throughput, if setting up a single lightpath between each source-destination pair, compared to that of a control-value. In this case the control-value is taken from the SNR-BA-random networks. This value indicates how much the physical properties have improved or worsened compared to those in the SNR-BA-random method. The average R_T values of all the networks are calculated and shown in Table 5.1. Values below 1 show a decrease and values above 1 an increase compared to the control-set.

$$R_T = \frac{\sum_z T_z}{\sum_z T_z^{control}} \tag{5.13}$$

It is clear that for all methods this ratio decreases compared to the control-set. In the worst case, throughput decreases on average by about 17% for the GA-DWC method. This means that the lightpaths setup by these topologies will have a worse achievable throughput compared to the control-set. However, if one can setup more lightpaths, then the overall maximum achievable throughput could still be higher.

Generally, the throughput penalty of increased path lengths is lower than the decrease in the number of allocatable lightpaths, as touched upon in section 5.3.1.1. This comes down to the inverse logarithmic scaling of throughput with increased transmission distance. This can be shown by calculating the average increase in path length compared to that of the control set, as de ned in Eq.(5.14), calculated for each topology and shown in Table 5.1.

$$R_P = \frac{\overline{L_P}}{\overline{L_P}^{control}} \tag{5.14}$$

This value shows that for the worst-case, GA-DWC on average increases the path length by 70%, however only demonstrates a 17% penalty within the average worst-case throughput, compared to the control-set. Therefore, a drastic increase in path lengths does not cause a proportionate response in achievable worst-case throughput.

The nal ratio investigated compares the maximum number of allocatable lightpaths (N_{LP}) from the ILP to that of the control-set. This ratio is de ned as in Eq.(5.15) and shown in Table 5.1.R_{LP} was used to understand the effect of number of allocatable lightpaths on the maximum achievable throughput. This property is only impacted by the structural properties of the network.

$$R_{LP} = \frac{N_{LP}}{N_{LP}^{control}} \tag{5.15}$$

Although the Greedy-DWC method does not have the best physical path ratios, it exhibits the highest number of allocated lightpaths with a 92% increase compared to the control-set. This in the end, results in the topologies with highest maximum achievable throughput, with an average increase in maximum achievable throughput of 82% compared to the control-set. However, the GA-DWC networks have the next largest average number of allocated lightpaths, with an increase of 88% over the control-set. However, due to a 70% increase in path length, the achievable SNR of these lightpaths degraded by about 17%. The networks designed by the GA-DWC-greedy method can accommodate 14% less lightpaths than the GA-DWC method, with a throughput degradation of 7% compared to the 17% of the GA-DWC networks exhibit. Due to this reduced SNR degradation, the GA-DWC-greedy networks have a maximum achievable throughput of 7% higher than that of the DA-DWC networks, despite the reduced number of allocated lightpaths. This

demonstrates that optimising networks only to increase the maximum number of lightpaths is not always the best option. However, balancing this design objective with that of the physical length of paths within the network is essential to optimising the maximum achievable throughput. A similar conclusion was observed in Chapter 3, where SNR-BA networks achieved higher maximum achievable throughput compared to ER and BA networks with about 9-11% less allocated lightpaths. However, this was an indirect result, in terms of that these structural and physical properties were not optimised in the SNR-BA networks to achieve this maximum achievable throughput. Therefore, the question remains, how can this relationship between structural and physical properties be included within the optimisation of PTD?

When using $= 0$ within the DWC, the optimisation has no concept of physical path length of the designed topology. Therefore, the GA-DWC method yields the lowest average DWC value of 0.33. This reduced DWC appears to lead to a lower number of hops within paths, however increased the physical path lengths of them. The proposal behind DWC was to optimise the structural and physical properties via path lengths. For the structural optimisation of the network, number of hops is to be reduced, with the idea that this allows lightpaths to use less wavelength resources and therefore allocate more lightpaths within the network. However, it seems this additional reduction of path hops did not translate into a larger number of setup lightpaths, at least not directly. The greedy-DWC method has a higher DWC value of 0.35, with a higher average number of lightpaths setup.

To conclude, optimising DWC with $= 0$, achieves on average a 63.1% increase in maximum achievable throughput compared to that of the control-set that does not take into account the objective of DWC. However, ignoring the physical path length in the pursuit of increased number of assigned lightpaths, can actually result in worse maximum achievable throughput values. Therefore, the question is, how does one balance this trade-off between number of setup lightpaths and the physical path lengths in the topology?

5.4.2 Demand Weighted Cost Minimisation $= 0{:}5$

In the previous section, it was shown that the optimisation of physical path lengths needs to be included in some manner, in addition to the optimisation of structural properties. This section investigates this by testing whether this is possible by altering the weighting the DWC to include the physical path lengths. The signi cance of this is, that instead of only weighting the paths by the number of hops, now the DWC includes an equal weighting of path lengths in terms of hops and distance. This weighting should balance the two objectives and make different edge choices. To do this, a new set of topologies is designed using the same methodology, however now

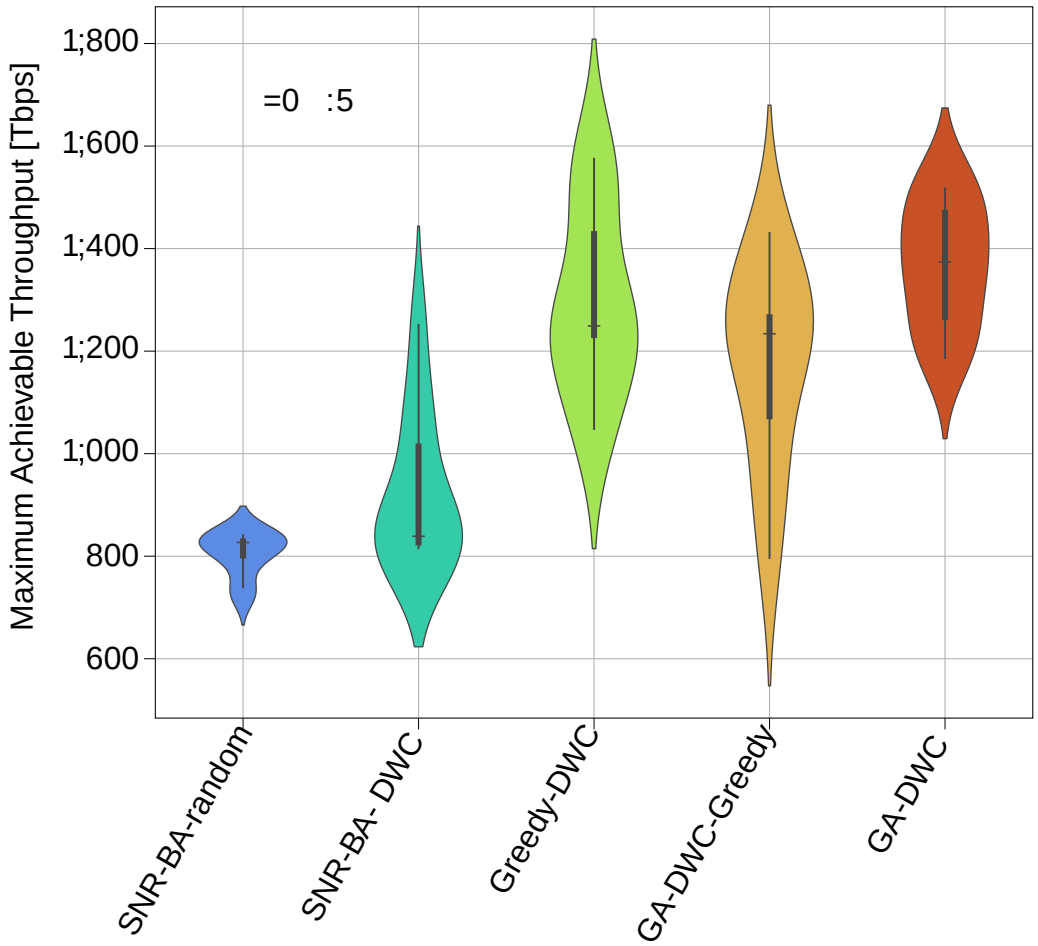


Fig. 5.3: Resultant maximum achievable throughput of networks designed using SNR-BA, Greedy and GAs to minimise the DWC at = 0 :5, compared to the control-set of SNR-BA-random.

with = 0 :5, to equally weight the path length by hops and distance. Again ten topologies are synthesised per design method and their maximum achievable throughput is calculated and plotted in Figure 5.3.

The Greedy-DWC method give the lowest average DWC values. However, the GA-DWC networks seem to outperform the other networks with 65% increased average maximum achievable throughput compared to the control-set. The DWC of these GA-DWC networks, however is the largest out of all of the designed topologies. Here R_T and R_P values of all the designed topologies lie close to 1, meaning that path lengths within the designed networks are largely unchanged from the control-set. Therefore, the physical properties of the newly designed networks do not have a signi cant impact on the nal maximum achievable throughput. The largest factor in nal performance, therefore, is that of N_{LP}, for which the GA-DWC has the highest number of allocated lightpaths, with 75% more lightpaths allocated than in the

control-set. GA-DWC, therefore, performs the best in terms of average maximum achievable throughput, due to the number of lightpaths it can assign, without degrading the SNR of these lightpaths much (1% degradation compared to control-set). However, the GA-DWC designed with $= 0:5$ on average achieve about 9% less maximum achievable throughput, compared to the best performing Greedy-DWC networks designed with $= 0:0$. This is because the objective of minimising path length in terms of hops does not directly allow for maximising the number of allocated lightpaths. Is there a way that one can directly optimise a topology and its resulting physical and structural properties?

To do this, the limiting cut method was re-evaluated to see whether it can be implemented in a greedy heuristic.

5.4.3 Limiting Cut Greedy Optimisation

Since the majority of additional throughput originates from topologies that can assign a larger number of lightpaths, the design objective should directly optimise this property. The maximum number of assignable lightpaths is restricted by a set of edges c_l, that separate the network into two sub-graphs with jN_1j F_j and jF_j nodes. The set of edges c_l that results in the maximum value of $\frac{jF_j(jN_1j\ F_j}{jc_lj}$ is called the limiting cut within a network, which also de nes the minimum number of wavelengths that are required to connect the network, given that a single lightpath is routed between all node-pairs, de ned in Eq.(4.2). This edge set, c_l, can be found exactly for networks of up to about 25 nodes, by iterating over all possible solutions. By dividing the total number of wavelengths available by the wavelength requirement estimates the number of assignable lightpaths for a uniform traf c demand. Multiplying this value by the worst-case average throughput gives a rough estimate as to how much throughput a network can achieve. This optimisation objective is de ned in Eq.(5.16). Obviously, this objective has one major drawback, namely that for larger networks an approximate algorithm needs to be used to estimate the edge set, c_l, which might not be the global maximum. In addition, for even 20 node networks, this objective is not realistic to implement in anything more than a greedy heuristic, which means that one forfeits on possible gains from more sophisticated combinatorial optimisation algorithms.

$$\underset{(u;v)2E}{\operatorname{argmax}} \quad \frac{B_T}{B_{CH} \quad d\max_{c_l 2 C_L}\left(\frac{jF_j(jN_1j\ F_j)}{jc_lj}\right)e} \quad \sum_{z2Z} 2B_{CH}\log_2(1 + SNR_z) \qquad (5.16)$$

An additional set of ten topologies with 20 nodes was designed using a greedy heuristic with the design objective de ned in Eq.(5.16). These networks are referred to as Greedy-Cut networks. Their maximum achievable throughput was calculated and

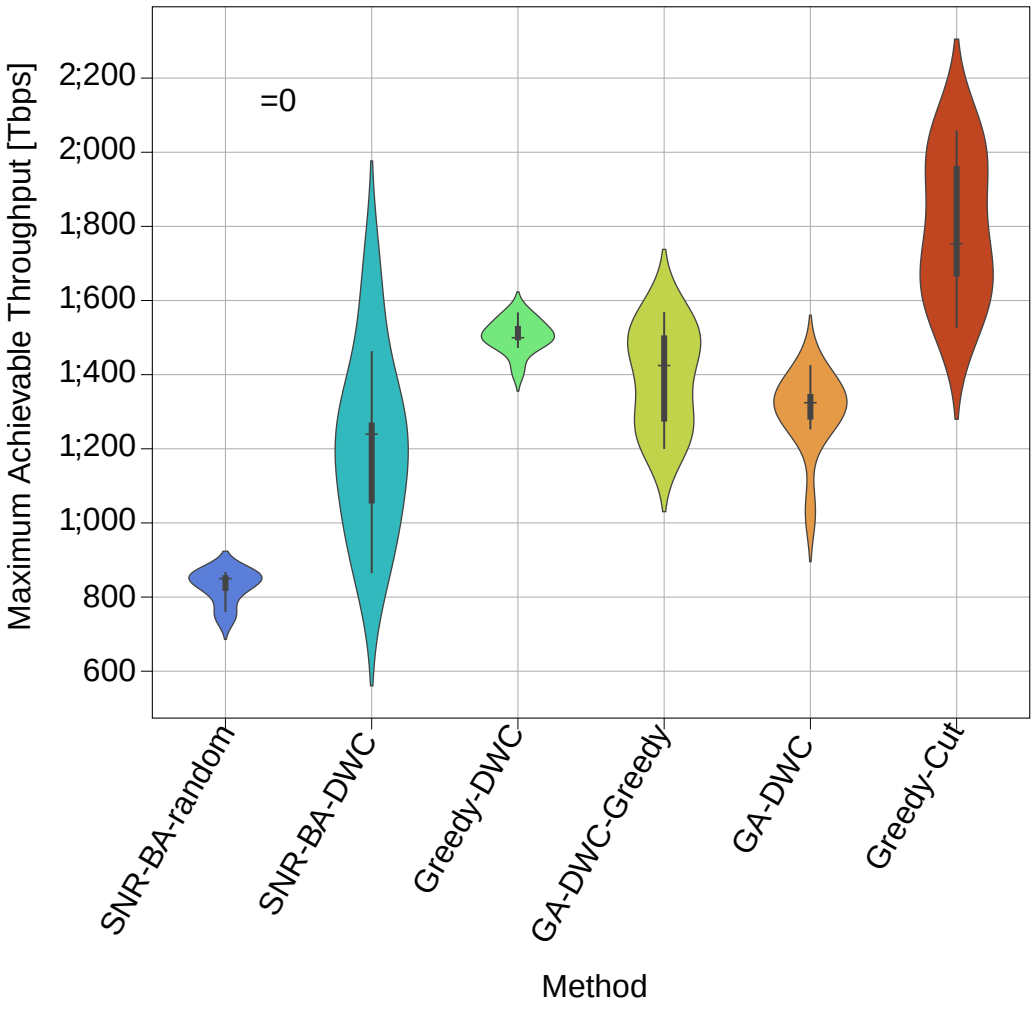


Fig. 5.4: Resultant maximum achievable throughput of networks designed using SNR-BA, Greedy and GAs to minimise the DWC at = 0 :0, compared to the control-set of SNR-BA-random and Greedy-Cut design method.

added to the other design methods, plotted in Figure 5.4.

From Figure 5.4, it can be seen that the greedy-cut method outperforms all the DWC-based optimisation methods in terms of maximum achievable throughput. On average the greedy-cut method more than doubles the maximum achievable throughput, with a 106% average maximum achievable throughput increase compared to the control-set and a 14% increase over the best performing DWC method. From Table 5.1, it is clear that this method has 114% more allocatable lightpaths ( ), with only an additional 7% larger average path lengths, making these networks the best performing topologies for this traf c demand. By optimising both the structural and the physical properties directly, rather than through a proxy such as DWC, one can see a huge improvement of maximum achievable throughput. This direct optimisation borrows concepts from both optimisations of DWC at = 0 and at = 0 :5, by

improving the maximum number of lightpaths assignable, as well as the resultant SNR of these lightpaths. In addition, the DWC of these networks are seen to be the highest of the design methods (excluding the control-set), showing that there is no direct correlation between DWC and maximum achievable throughput. Within the DWC metric, average physical path lengths directly impact maximum achievable throughput, as was seen with R_P, however minimising the number of hops within paths does not have a direct optimisation effect on the number of setup lightpaths. The greedy-cut method, however, includes a direct optimisation of the number of allocatable lightpaths and the throughput of these lightpaths, thus achieving the highest maximum achievable throughput overall. This additionally demonstrates, that accurate design objectives are important and can make a signi cant difference, in contrast to solely focussing on improved combinatorial optimisation. To demonstrate how this objective can be used in more advanced optimisation, the following section investigates the inclusion of the limiting cut objective in a deep reinforcement learning algorithm.

5.4.4 Advanced PTD Optimisation

The greedy-cut method was shown to be a direct optimisation compared to the DWC-based optimisation methods. This meant that due to the higher accuracy of the objective function - by using the limiting cut and GN-model - the greedy-cut method was able to improve the maximium achievable throughput by 14% compared to the best DWC method. Another bene t of this direct optimisation, is that the optimisation method is essentially not "fooled" and can distinguish the maximum achievable throughput of potential solutions accurately. A greedy heuristic performs well in terms of computational complexity, however it does not include the combinatorial optimisation of all the edges together. It only picks the best edge choice irrespectively of other edges. This did not seem to matter much within the DWC optimisation, since the DWC was indirectly linked to maximum achievable throughput. This meant that the GA could achieve a better DWC value, however not a better maximum achievable throughput. To demonstrate how the more accurate limiting cut objective function can enable more complex optimisation, deep reinforcement learning is applied to the PTD problem using the limiting cut objective described in Eq.(4.3).

5.4.4.1 Deep Reinforcement Learning

Deep reinforcement learning is a branch of machine learning that has adapted deep learning architectures to act as estimators for speci c functions used in traditional reinforcement learning [15]. Much research has applied deep reinforcement learning to the RWA problem with little or overstated success. However, more recently this has also been applied to topology design of general networks [160, 161]. Reinforcement

learning uses markov decision processes to model a sequential decision process, such as the PTD problem. A markov decision process uses a state (S_{rl}), an action (A_{rl}) and a reward (R_{rl}) to model a sequential decision process. Within topology design the state can be seen as the Graph (G_t) at a particular time step, an action can be modelled as a node to pick or an edge to pick (where two nodes give an edge to add). The reward is the design goal, in this case the maximum achievable throughput, modelled by the limiting cut objective function. In deep reinforcement learning, an agent (a neural network architecture) interacts with an environment by choosing an action from a set of legal actions to perform. These actions change the state of the environment and potentially gives a reward signal. However, the agent might interact many times with the environment before receiving a reward signal. The goal of the agent is to interact with the environment by choosing actions in a way that when the interactions end, the overall reward is maximised, in this case the maximum achievable throughput. At each time step, this is referred to as the discounted future return at time t and is de ned in

$$R_t = \sum_{t'=t}^{T_{rl}} {}_{rl}\, r_{t'} \tag{5.17}$$

where T_{rl} is the nal timestep, $r_{t'}$ is the reward at time step t' and ${}_{rl}$ is the discount factor. The discount factor is used to ensure a nite sequence is used and is always smaller than one. The goal is to maximise Eq.(5.17) over all time steps. To do this, actions need to be selected accordingly. These actions need to be selected by the future expected value (reward) that they will deliver to the environment for any possible policy in the future. This is normally estimated by an action value function $Q(a_t; s_t)$. $Q(a_t; s_t)$ measures the future expected return of an action taken at time t (a_t) given a state (s_t). The optimal Q follows the bellman equation de ned as

$$\mathop{E}_{s'} \; r + {}_{rl} \max_{a'} (Q(s', a') | s; a \tag{5.18}$$

where a', s' represent the actions and state following a and s in this case. The idea is that if Q is known for the sequence s' at the next time-step for all possible actions a', then an optimal strategy could be to select the action a' which maximises the expected value of $r + {}_{rl} Q(s', a')$. In deep reinforcement learning this Q-function is approximated using deep learning architectures, generally ANNs. In case of graph learning it makes sense to use graph neural networks. Therefore, we repurpose the MPNN studied in Chapter 4 to learn an approximation to the Q-function.

The MPNN parameters are referred to as ^{MPNN}. To train the MPNN the Q-function is evaluated using $Q(s; a; {}^{MPNN})$ and compared against a target network. This target network - with parameters - is a copy of the MPNN at an older training iteration and is updated throughout the training. This target network is required, since

if the same network is used to measure the target value, then on each training iteration the target would change due to parameter update. This would cause the training to be extremely unstable. Therefore, the final loss function at iteration i is defined as

$$L_i(\theta_i) = \left[r + \gamma_{rl} \max_{a'}(Q(s', a', \theta_i) - Q(s, a, \theta_i^{MPNN}) \right]^2 \quad (5.19)$$

The simplest of deep reinforcement learning architectures, came in the form of deep q-learning (DQN), successfully demonstrated by Mnih et al. [162], where an agent learnt to play a set of Atari games. DQN is an off-policy algorithm, meaning that the experiences used for training are not directly generated with the up-to-date policy. The experiences are sampled from a buffer of pre-determined length, this is referred to as experience replay.

The algorithm followed for training is described in algorithm 6. Initially experiences are collected by an epsilon-greedy method, where a parameter ϵ_{rl} is set at the beginning (normally close to one) and random actions are sampled according to this. If not random, an action is sampled using the greedy argmax of the MPNN θ^{MPNN}. ϵ_{rl} is decayed over a certain number of exploritory steps within the algorithm. The actions chosen are executed in an environment $Env(G_t, a_t)$, which returns the new state of the graph G_{t+1} and a reward r_t. This action (a_t) taken on graph G_t, which results in the reward r_t and altered graph G_{t+1} is saved as an experience in the replay buffer (β_R). When a training step is taken, a mini-batch is sampled from the replay buffer. For each sample (j) in this mini-batch, the target y_j and the loss $L_j(\theta^{MPNN})$ are calculated and used to perform a gradient-descent step. Then every C_t steps the target network (θ) is reset by the up-to-date Q-network θ^{MPNN}.

5.4.4.2 PTD States, Actions, Rewards and Q-network

As described in the previous section, the Q-network is implemented by the same MPNN as described in Chapter 4. The implementation of this MPNN is mostly identical, except for the readout layer. Within Chapter 4, the MPNN solved a regression problem, meaning that the final layer outputted a single value, the maximum achievable throughput. The MPNN in this case, however, is estimating the Action Value function, which estimates the future discounted reward of an action. Therefore, it outputs the future discounted reward for each action that is possible to take, from which the argmax is taken to choose the next action in the sequence. The output size of the readout function then goes from a single value to that of size $|A_{rl}|$.

The state used by the environment is that of the graph (G_t), which uses the same node features - nodal degree and traffic - as in Chapter 4. Actions follow the same representation as in [160], where each action represents a node in the graph, where two separate actions give an edge which is added to the graph. At the end of an episode (after

Algorithm 6: DQN training algorithm
Input: T_z^C, G
Output: G_D

```
Initialise replay buffer B_R;
Initialise MPNN θ^MPNN;
Initialise Target MPNN θ⁻ ← θ^MPNN;
for t ∈ T_G do
    x_eg ~ Bern(ε_rl);
    if x_eg then
        a_t = random;
    else
        a_t = argmax_{a∈A} Q(G_t, a; θ^MPNN);
    end
    r_t, G_{t+1} = E_rl(G_t, a_t);
    Add a_t, G_t, G_{t+1}, r_t into B_R;
    G_t = G_{t+1};
    Sample random minibatch from B_R;
    y_j = { r_j                                   if episode is finished
          { r_j + ε_rl max_{a'}(Q(s_{t+1}, a'; θ⁻))   otherwise ;
    Perform gradient descent with loss
    L_j(θ^MPNN) = (r + ε_rl max_{a'}(Q(s', a'; θ⁻)) − Q(s, a; θ^MPNN))²;
    Every C_t steps reset θ⁻ = θ^MPNN;
end
end
```

all intended edge additions have been made), a reward is given by the environment. This reward is the the objective function described by Eq.(5.16).

The Q-network was trained by adding 8 edges into the network, after which the graph is reset again. At the beginning of each episode a random graph from the 10 20 node networks previously used in this chapter is chosen to be designed. Initially, 50,000 global steps are taken and stored in the replay buffer without starting the training. After this the training is started, where on every forth iteration a training iteration is performed. A training iteration samples a mini-batch of 32 samples, calculates the loss and performs a gradient descent step. In total approximately 15 million steps are taken to train the agent on a Nvidia A100 GPU.

The final agent is then saved and evaluated. The evaluation is then done by a process called beam-search, where instead of taking the argmax at each timestep, the top three actions are followed for the first 8 actions (4 edges), after which the final 8 actions are then chosen via argmax. This gives 6561 solutions, of which the best is

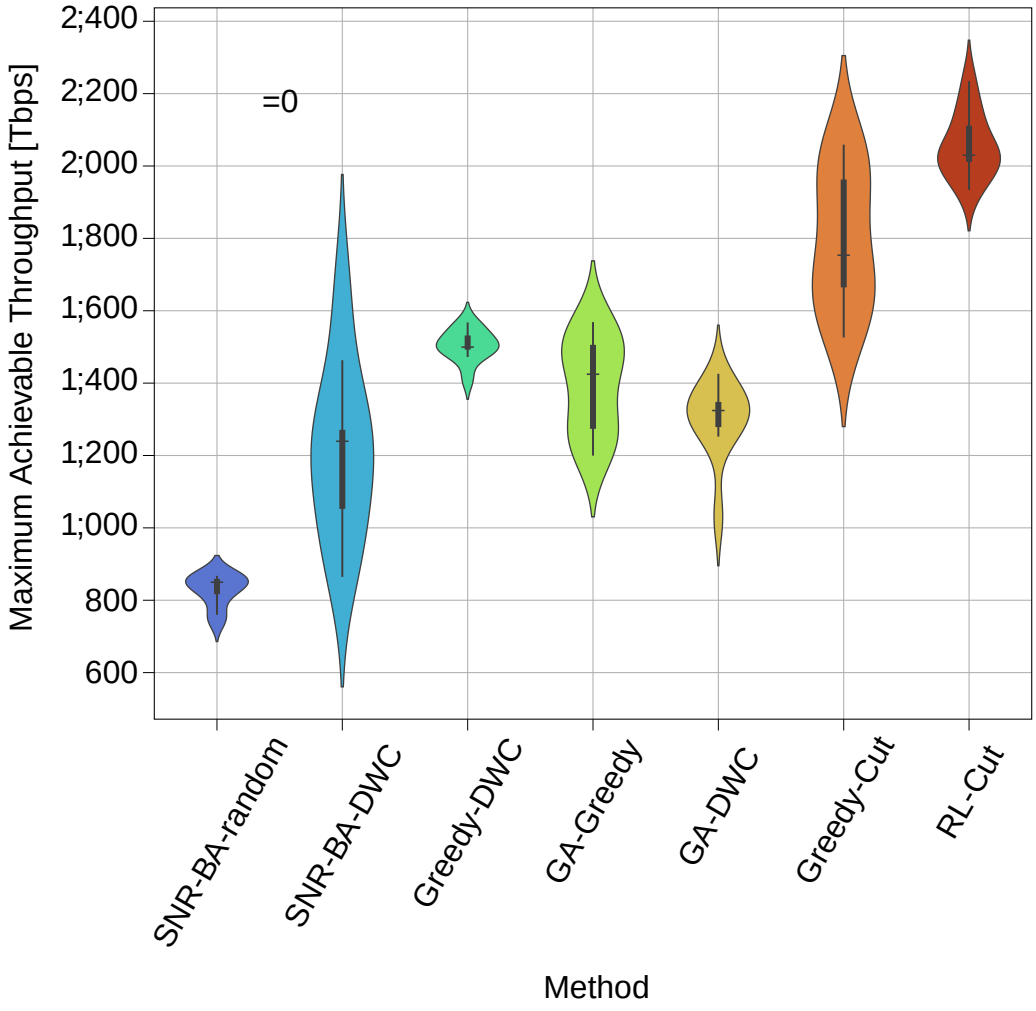


Fig. 5.5: Resultant maximum achievable throughput of networks designed using SNR-BA, Greedy, GAs, Greedy and deep reinforcement learning (RL) cut methods to maximise achievable throughput at = 0 :0, compared to the control-set of SNR-BA-random.

then selected and the ILP evaluated for maximum achievable throughput and plotted in gure 5.5. Here it can be clearly seen that the RL-agent is able to improve the maximum achievable throughput on average by 14% compared to the Greedy-Cut method. Additionally, the worst case is improved by 27%. From table 5.1, one can see that the RL-Cut networks have an all-to-all throughput ratio ( ) that is 8% less than that of the SNR-BA-random networks, caused by 29% longer paths. However, due to a 168% increase in actual lightpaths allocated, there is a 148% increase in maximum achievable throughput compared to the SNR-BA-random networks.

This further improvement demonstrates how this objective function enables more advanced optimisation to give better solutions and novel designs.

The investigated design problems all are performed using uniform traf c, meaning

that all node-pairs request the same number of lightpaths between them. To understand better the performance under different traf c demands, the following section investigates topology design under skewed traf c.

5.4.5 Skewed Traf c Analysis

As optical networks inter-connect many larger cities and countries, the nodes generally service varying amounts of traf c depending on population centres, data centre size and placements, amongst other variables [120]. The traf c itself is signi cant to the PTD, because different traf c will cause different bottle-necks in the network and affect the maximum achievable throughput. Therefore, investigating a systematically skewed traf c is bene cial to future network design, by ensuring robustness of design methods. To skew the traf c demand $\overline{T_z^C}$, a simple scheme is developed. At random, two node-pairs are chosen, without replacement, until all have been chosen. For each of these pairs (z_1, z_2) a pre-de ned skewed amount $_T()$ of traf c is taken from one node-pair and given to the other, as de ned in Eq.(5.20).

$$\overline{T_{z_1}^C} = \overline{T_{z_1}^C} + {}_T \quad \overline{T_{z_2}^C} \tag{5.20}$$

This keeps the total traf c the same, however skews it according to $_T$. As traf c values are exchanged among all node-pairs, the maximum skew of $_T = 1$, results in all the traf c being routed between only half of the node-pairs. Using $_T = [0:2; 1:0; 0:4]$, three sets of ten topologies are designed for each design algorithm, except for the RL-Cut method. This was excluded since the agent was not trained over varying traf c demands, which is left for future work. In addition, the limiting cut was altered to accommodate for non-uniform traf c and the average throughput T_2 sum was weighted by $\overline{T_z^C}$ to give a higher weighting to more frequently requested node-pairs. The maximum achievable throughput for all these topologies was then calculated and presented in Table 5.2.

Method	DWC	LP allocated	Throughput [Tbps]	R_T	R_P	R_{LP}		τ
SNR-BA-random	4.63	1510	832.47	1.00	1.00	1.00	0.0	0.2
SNR-BA-DWC	0.39	2042	1109.97	0.97	1.05	1.35	0.0	0.2
Greedy-DWC	0.35	2850	1466.14	0.91	1.31	1.88	0.0	0.2
GA-DWC	0.33	2793	1279.13	0.82	1.79	1.84	0.0	0.2
Greedy-Cut	0.40	3249	1754.14	0.97	1.06	2.15	0.0	0.2
SNR-BA-random	4.05	1463	804.08	1.00	1.00	1.00	0.0	0.6
SNR-BA-DWC	0.40	1928	1043.97	0.98	1.03	1.31	0.0	0.6
Greedy-DWC	0.35	2755	1417.39	0.92	1.26	1.88	0.0	0.6
GA-DWC	0.33	2726	1266.38	0.84	1.67	1.86	0.0	0.6
Greedy-Cut	0.40	3296	1769.64	0.97	1.05	2.25	0.0	0.6
SNR-BA-random	4.33	1539	850.13	1.00	1.00	1.00	0.0	1.0
SNR-BA-DWC	0.40	2023	1097.41	0.97	1.03	1.31	0.0	1.0
Greedy-DWC	0.35	2622	1348.85	0.92	1.26	1.70	0.0	1.0
GA-DWC	0.33	2555	1164.94	0.82	1.67	1.66	0.0	1.0
Greedy-Cut	0.40	3220	1746.51	0.97	1.05	2.09	0.0	1.0

Table 5.2: Average demand weighted cost, lightpaths allocated, maximum achievable throughput, R_T, R_P, R_{LP}, and traffic skew τ for designed topologies compared against the control-set.

Table 5.2 shows that for each skew value ( ), the Greedy-Cut method outperforms all other methods in terms of number of allocated lightpaths and maximum achievable throughput, on average improving upon the control-set by 112% over all skew values. The Greedy-Cut method, in terms of maximum achievable throughput, outperforms the best performing DWC method on average by 25%.

It is clear that all the designed topologies have similar DWC values, however the performance hierarchy is not predicted well by DWC amongst the designed topologies again. This was highlighted in section 5.4.3 and is also the result for the case of skewed traffic. DWC causes a shift in maximum achievable throughput compared to the control-set, however no direct correlation between maximum achievable throughput and DWC can be shown. Therefore, only through the direct optimisation of cutsets and SNR of paths present in topologies can one fully optimise the maximum achievable throughput of these networks. This is difficult to implement for larger networks ( ), however. This has recently started to be investigated by authors in [19].

A topological analysis was performed to further understand the well-performing topologies and is described in the next section. Its goal is to understand the topological attributes of the best-performing topologies.

5.4.6 Topology Structural and Physical Properties Analysis

The degree distributions were calculated for the designed networks for all values of traffic skew ( ) and plotted in Figure 5.6. This was done to investigate the structural properties of these networks. From Figure 5.6 one can see that the degree distribution resulting from the Greedy-Cut networks is more akin to that of the SNR-BA-random networks. The SNR-BA-random and Greedy-Cut networks generally have fewer high-degree nodes than the other methods, with a maximum degree of about 7. Whereas Greedy-DWC, GA-DWC and SNR-BA-DWC have more high-degree nodes present, with maximum degrees of 10 respectively. The Greedy-Cut networks have a higher number of allocated lightpaths, which means they have a lower wavelength requirement compared to the other designed topologies, which indicates that these networks are better connected for the traffic that they serve.

Between uniform traffic at a value of = 0:0 (Figure 5.6a) and the skewed-traffic patterns (Figure 5.6b-d), one can notice that the degree distribution of the Greedy-Cut networks is skewed more towards higher degree nodes. This shows that at uniform traffic it seems that it is better to equally distribute edges across nodes, however with higher skew values, there is benefit in skewing the degree distribution towards certain nodes. In the skew evaluated of = 1:0, traffic is only requested between half the node-pairs in the network. Therefore, concentrating edges on specific nodes makes

sense here, i.e. more higher degree nodes. In addition, the Greedy-Cut networks have a maximum degree of 7, whereas the other designed networks have maximum degrees of around 7-10. In chapter 3 these types of highly connected hubs, attributed to the BA networks, were shown to give better wavelength requirement performance. It is clear from comparing the degree distributions in Figure 5.6 and the maximum achievable throughput performance of the GA-DWC and Greedy-Cut methods in tables 5.1 and 5.2, that highly connected hubs (for example in GA-DWC) can give improved wavelength requirement performance compared to the control-set, however it is not a direct optimisation of the property itself. In other words, aiming for properties such as minimal path lengths, minimal DWC, or larger number of highly connected hubs, wavelength requirement performance improves, however these properties do not directly optimise wavelength requirements. Greedy-Cut method does not aim for these intermediate objectives, however optimises networks along the limiting cut, which is the true bottle-neck in performance and leads to a direct way to maximise achievable throughput.

Other structural metrics were further investigated, such as the mean values for diameter, algebraic connectivity (smallest eigenvalue), spectral radius (largest eigenvalue), second largest eigenvalue, average clustering coef cient, average edge disjoint paths (EDP) between node-pairs. These were calculated and listed in Table 5.3. Here all eigenvalues were calculated from the normalised Laplacian. The clustering coef cient is a property of the network that measures the probability of whether nodes will share the same neighbours, whilst edge-disjoint paths are paths that exist in the network without edge overlap. The latter is particularly important within optical networks, as edge-disjoint paths can route multiple lightpaths using the same wavelengths. However, when paths reuse edges within the network, two lightpaths require different wavelengths, therefore using more resources.

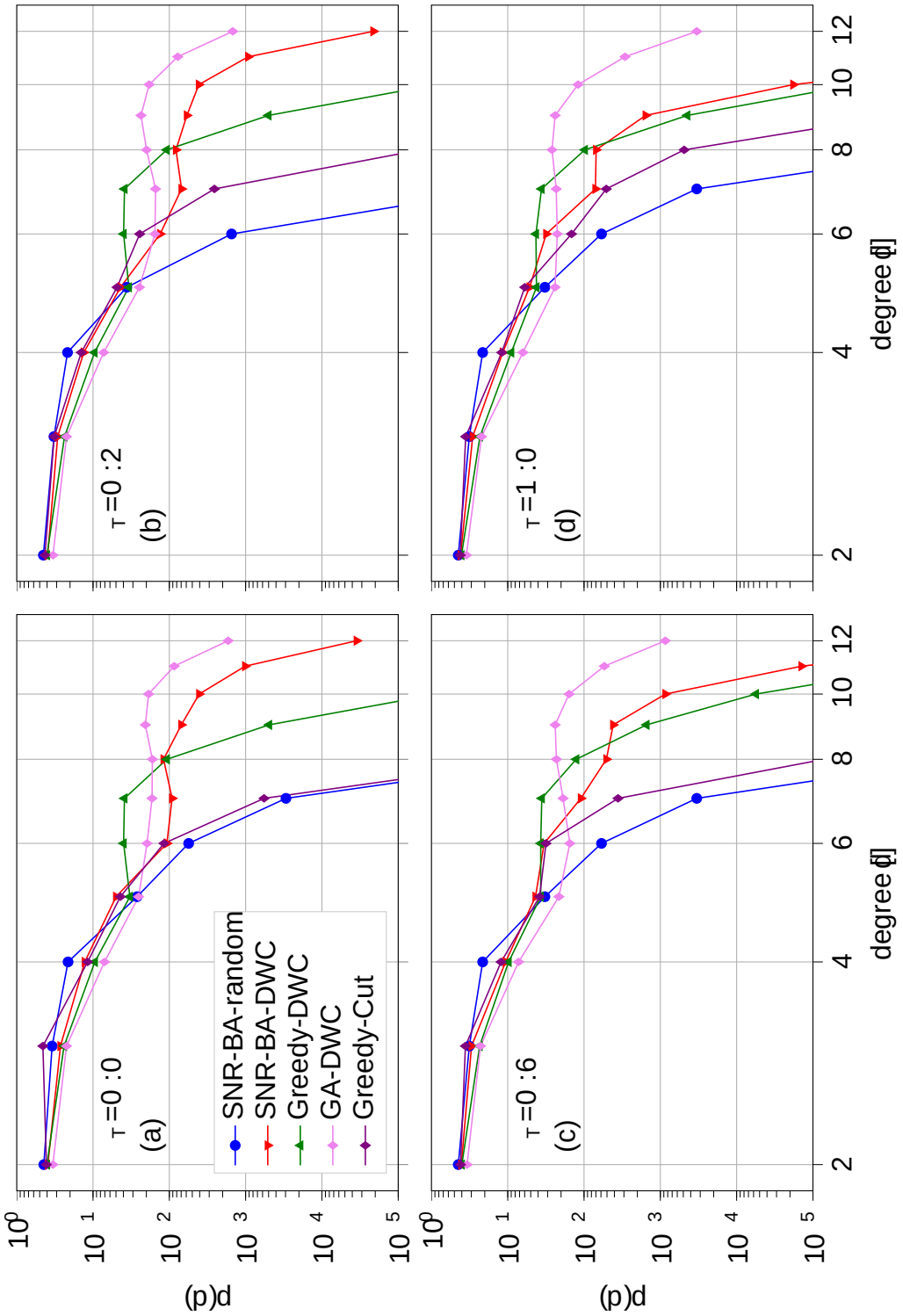


Fig. 5.6: Degree distributions of networks designed using SNR-BA-random, SNR-BA-DWC, Greedy-DWC, GA-DWC and Greedy-Cut methods. Each Figure presents a different traf c skew χ.

When analysing the average diameters of the networks in Table 5.3, it is clear that the designed topologies have 2-4 hops smaller values of the diameter than that of the control-set. The Greedy-Cut method does not have the lowest average diameter amongst the design methods though, indicating that diameter - like DWC - generally improves wavelength requirements, however does not directly determine it.

The algebraic connectivity has been investigated multiple times for correlating well to wavelength requirements and proposed as a optimisation function before. Higher values are associated with more resilient and higher throughput networks [26, 25, 21]. From Table 5.3 it is clear that the designed topologies have higher algebraic connectivities than the control-set, however the Greedy-Cut method has one of the lowest algebraic connectivities. This is surprising since the Greedy-Cut method has the lowest wavelength requirements in each of the designed topologies. Again this shows that the algebraic connectivity - like DWC - can give networks with lower wavelength requirements, however it does not directly optimise the wavelength requirements as discussed when investigating the degree distributions.

The Greedy-Cut method displays similar low clustering coef cients as the other best-performing DWC-design methods, however it is evident that the average number of edge disjoint paths per node-pair is higher than that of all other networks for each traf c skew. This means that, on average, there are more paths that do not share edges, therefore multiple lightpaths can use these different edge disjoint paths, without using alternative wavelengths. This means that more lightpaths can be routed using the same resources compared to other networks.

Lastly, the spectral radius (and EDP) is one of the only values analysed (except for the number of edge disjoint paths) that consistently show a heightened value within the structural analysis. The spectral radius is associated with the robustness of networks, however the spectral radius calculated from the adjacency matrix has been shown to give an upper bound on the chromatic number of graphs, which is equivalent to the wavelength requirements in optical networks given uniform traf c [163]. The spectral radius calculated from the distance Laplacian has previously been shown to correlate well to maximum achievable throughput, however the exact relationship between the spectral radius and the maximum achievable throughput is not known. The Greedy-Cut networks tend to have higher spectral radii, compared to the other designed networks, whilst also having improved maximum achievable throughput. This again is inline with previous research that determined this to be an important property within chromatic numbers, however also the maximum achievable throughput [21].

The structural investigation of the designed topologies has revealed interesting results. Namely that previous network design principles of algebraic connectivity and highly-connected hubs hold true, however are not optimal in terms of maximum achievable throughput.

Method	Diameter	λ_1	λ_n	λ_{n-1}	Clustering	EDP	T
SNR-BA-random	8.2	0.038	1.89	1.81	0.35	2.05	0.0
SNR-BA-DWC	5.7	0.122	1.92	1.85	0.15	2.14	0.0
Greedy-DWC	5.0	0.207	1.90	1.80	0.02	2.15	0.0
GA-DWC	4.3	0.169	1.90	1.84	0.07	2.06	0.0
Greedy-Cut	5.3	0.121	1.94	1.88	0.05	2.36	0.0
SNR-BA-random	8.5	0.043	1.89	1.80	0.33	2.07	0.2
SNR-BA-DWC	6.1	0.112	1.90	1.85	0.17	2.13	0.2
Greedy-DWC	5.0	0.211	1.89	1.79	0.02	2.15	0.2
GA-DWC	4.1	0.158	1.90	1.85	0.07	2.06	0.2
Greedy-Cut	5.1	0.135	1.93	1.88	0.07	2.28	0.2
SNR-BA-random	8.0	0.043	1.89	1.79	0.34	2.06	0.6
SNR-BA-DWC	6.0	0.103	1.90	1.86	0.16	2.12	0.6
Greedy-DWC	4.8	0.188	1.88	1.81	0.04	2.15	0.6
GA-DWC	4.1	0.190	1.90	1.82	0.07	2.07	0.6
Greedy-Cut	5.0	0.149	1.94	1.86	0.07	2.30	0.6
SNR-BA-random	8.2	0.042	1.90	1.81	0.33	2.06	1.0
SNR-BA-DWC	6.0	0.087	1.89	1.85	0.20	2.12	1.0
Greedy-DWC	4.9	0.191	1.87	1.81	0.03	2.15	1.0
GA-DWC	4.5	0.173	1.90	1.83	0.08	2.07	1.0
Greedy-Cut	5.1	0.119	1.93	1.87	0.10	2.30	1.0

Table 5.3: Mean diameter, algebraic connectivity (λ_1), spectral radius (λ_n), second largest eigenvalue (λ_{n-1}), clustering, edge-disjoint paths per node-pair (EDP) of designed networks for varying values of traffic skew (T).

Highly connected hubs were shown to give improved wavelength requirement performance in BA-generated networks in Chapter 3. Similar highly-connected hubs were constructed in the GA-DWC networks, which showed that they improved wavelength requirements with respect to the control-set. However, it was shown that in the case of greedy-DWC that with a more uniform degree distribution compared to that of the GA-DWC network and a smaller maximum degree within networks, they can achieve better wavelength requirement performance. This is consistent with the findings in Chapter 3 that ER networks have slightly improved wavelength requirements, compared to BA and SNR-BA networks. ER networks give degree distributions that are normally distributed, i.e. more uniform which was seen in figure 3.4 in Chapter 3.

In addition to this, lower diameter and higher algebraic connectivities have been connected to lower wavelength requirements in networks [26]. All designed networks show higher algebraic connectivity and lower diameters compared to the control-set. However, the best performing Greedy-Cut networks do not exhibit the lowest diameters or highest algebraic connectivities. The only metrics which consistently were found to be the largest in networks were the spectral radii and the average number of edge disjoint paths of the Greedy-Cut networks. From this one can

conclude that higher algebraic connectivity, highly-connected hubs, smaller diameters and lower DWC are good design strategies, however they do not lead to optimal networks in terms of maximum achievable throughput.

5.5 Summary

This chapter investigated the problem of including maximum achievable throughput as an objective within the PTD problem. This was proposed by minimising the demand weighted cost (DWC) of the network, shown to have a strong correlation to maximum achievable throughput in Chapter 4. DWC was included as an optimisation objective within the SNR-BA generative graph model (SNR-BA-DWC), a greedy heuristic (Greedy-DWC) and two genetic algorithm variants (GA-DWC and GA-Greedy). It was shown that by minimising the DWC of the networks an average 63% increase in maximum achievable throughput was achieved compared to a control-set. However, the minimum DWC networks did not correspond to the highest maximum achievable throughput networks.

Using three ratios, the effect of structural (R_{LP}) and physical (R_T and R_{LP}) properties on the maximum achievable throughput was demonstrated. The design objective of DWC was able to account for the physical aspects directly (physical path length), however the structural optimisation of minimising path hops showed to not lead directly to a higher number of allocatable lightpaths. This meant that DWC was not a direct maximisation of achievable throughput.

Therefore, to improve on DWC, a new objective, based on the limiting-cut method was de ned in Eq.(5.16) and implemented in a greedy heuristic (Greedy-Cut). This method increased the number of allocated lightpaths by 114%, whilst increasing the average path length by just 7% and therefore achieving on average a maximum achievable throughput increase of 106% compared to the control-set. The limiting cut combined with the physical layer model demonstrated the direct optimisation of maximum achievable throughput, however currently not scalable to larger networks ($|N| \leq 25$). Although currently actively being researched by other authors [19].

Finally, the topological makeup of the designed networks was analysed by investigating: the degree distributions, diameter, algebraic connectivity, clustering coef cient, spectral radius, second largest eigenvalue, average clustering coef cient and average edge disjoint paths. The designed networks showed that previous objectives of including highly connected hubs, maximising algebraic connectivity, minimising diameters and minimising DWC all improved the maximum achievable throughput compared to the control-set. However, they did not show a direct maximisation of maximum achievable throughput. The problem remains however that these intermediary objective functions and design-principles do not lead to optimal

networks. Even if optimal combinatorial optimisation were to take place, the optimisation would be 'fooled' by objective functions used. Therefore, future focus should lie on the optimisation of graph cuts in the network, as the optimisation will always be limited by the objective function it is tasked to optimise. Better combinatorial optimisation including the cutset is a promising future research direction to further maximising achievable throughput of optical networks.

Chapter 6

Conclusions and Future Work

The overarching goal of this thesis was to directly maximise achievable throughput within optical core networks. Within the context of maximising the achievable throughput of optical networks, there were three distinct challenges that were explored. Firstly, the impact of both structural and physical properties on the maximum achievable throughput, was not well-known before this PhD research. Secondly, the inclusion of maximum achievable throughput as an objective function is dif cult, as its calculation requires an optimal solution to the RWA problem (NP-hard). Therefore, solving this problem in linear time is necessary to include it in the PTD problem. Finally, the PTD problem itself is NP-hard and the combination of search-strategy and objective estimation impacts the nal maximum achievable throughput of network solutions. Therefore, the nal goal of this research was to introduce intelligence into the PTD problem by maximising achievable throughput in optical core networks, in a way that is adaptable to distance and traf c.

Chapter 3 initially investigated different generative graph models that have previously been used to model optical networks and their shortcomings, whilst also presenting a new generative graph model - the SNR-BA model - that more accurately represents real networks. The SNR-BA model demonstrated this by having similar degree, diameter and spectral distributions to that of a set of real optical networks. The results showed that this model can be used for future optical network studies rather than single or a couple of real network topologies. Demonstrating performance of algorithmic solutions to the RWA and other operational problems on a set of topologies is important to distinguish between algorithmic performance and topological artefacts. A recent surge in the application of machine learning to operational problems within optical networks, has made this even more important. This is because these works only test on singular or few topologies, however to demonstrate the bene t of machine learning solutions this is not suf cient [58, 61, 17, 67]. This additionally motivated the synthesis of an open-source dataset (Topology Bench) for optical networking research, which includes an exhaustive set of 105 real optical networks and 270,000 SNR-BA networks [113].

Prior to the research in this thesis, the impact of structural and physical properties

on the maximum achievable throughput of optical networks was not well understood. It had been researched for particular topologies and from a data centre perspective where physical properties are neglected. Yet, understanding this was important, as it governs the maximisation of achievable throughput, impacted by both structural and physical properties, rather than the minimisation of wavelength requirements, dependent only on the structural properties of graphs. Only by understanding exactly how to manipulate the structural and physical properties of optical networks, can one maximise the achievable throughput within them. Although maximising throughput is not the only objective of optical network operators, understanding how these properties impact the nal throughput can help operators make choices that minimise the impact on maximum achievable throughput on the network.

In Chapter 3, it was shown that the SNR-BA model could be used to investigate the impact of structural and physical properties on the maximum achievable throughput of optical networks. The investigation revealed that physical properties, in addition to the structural properties of the network, impact the maximum achievable throughput. Although the SNR-BA networks on average allocated 8-11% fewer lightpaths and had 30-108% higher wavelength requirements than the ER and BA networks, they were able to achieve between 30-59% higher maximum achievable throughput. This was due to the ER and BA networks having 95-215% physically longer lightpaths than the SNR-BA networks. These results demonstrated that both structural and physical properties impact the maximum achievable throughput of optical networks and need to be included in network design. Although this is intuitive, given the extensive modelling of point-to-point optical transmission, this was not published or formally included in research prior to this investigation. Therefore, to include intelligence in the design of optical networks, it is important to take both the structural and physical properties into account, as to maximise the achievable throughput and therefore the performance of the optimisation. In addition, although operators rarely add additional edges into the network, often capacity is expanded over existing edges within the network. To improve the structural and physical properties of the network, operators need to know which edge capabilities to expand, vital for cost-effective growth of optical networks. The research in this thesis can be used to understand exactly this, which edges to add/upgrade.

Given the computational complexity of calculating the maximum achievable throughput of optical networks - discussed at the beginning of Chapter 4 - the next step was to explore methods for calculating this property with reduced computational complexity. This was necessary to include the maximum achievable throughput in the combinatorial optimisation in the network design stage. To overcome the computational complexity of the RWA problem, in section 4.1, simpli ed upper-bounds on maximum achievable throughput, which relied on a linear

programming optimisation, were derived and investigated. The linear program showed that it was able to calculate the maximum number of allocatable lightpaths - within 10% of the ILP solutions - for the majority of networks (95%). Ideas from [118, 122, 119] led to the derivation of this linear programming model, however the previous work was split between maximum achievable throughput in data centres and optical networks. Within the optical networks research, this had only been investigated from an operational point-of-view, where particular topologies were investigated, rather than a range of topologies. Therefore, the quanti cation of how accurate this LP estimation is to the ILP formulation, had never been investigated before. However, although incredibly accurate, the linear program still was not fast enough (within ms) to run for large-scale optical network design to maximise achievable throughput. This upper-bound, however, is worth expanding to include the GN-model and can be used for polynomial-time maximum achievable throughput calculation, in situations where the ILP formulation is infeasible. This is required, for example, the accurate study of how different traf c distributions affect maximum achievable throughput of optical networks in the future.

Previously, message passing neural networks (MPNN) were shown to be successful in modelling complex molecule properties in quantum chemistry [22], which inspired the rst application of message passing neural networks for modelling maximum achievable throughput in optical core networks within this PhD research. The three MPNN models - trained on networks between 10-100 nodes - achieved on average predictions within 8% of the calculated maximum achievable throughput labels, with high linear correlations (R^2) between 0.97-0.98. In addition, the inference of the model ran within ms, with huge parallelisation possible on GPUs. This demonstrated a huge 5 orders of magnitude savings over that of the ILP. However, when tested on networks with structures differing to those trained upon, the MPNN did not retain its high accuracy and due to this poor generalisation it was dif cult to include in the network design process. However, the MPNN model showed huge potential in modelling a nonlinear property (maximum achievable throughput) of a network and was the rst application of MPNNs for modelling maximum achievable throughput in optical core networks. This model is worth expanding on in the future and can be used to evaluate network scenarios and dynamic changes in the network quickly (within ms) and accurately (within 10%). In addition, although work in [135] was developed around the same time with a similar application, the case of throughput calculation in optical core networks, compared to data centre networks, requires much more complex optical channel modelling and therefore is in itself a novel application of MPNNs for maximum achievable throughput estimation.

Finally, the focus moved to the analytical metrics that correlate with maximum achievable throughput (section 4.3), as was done for wavelength requirements in the

past [93, 26]. The DWC was derived to include both structural and physical properties of paths and a weighting according to the traf c distribution. For both ILP and FF-kSP data, the DWC was shown to achieve high correlations between 0.87 and 0.97, with computational complexity that grows linearly with network size. This metric can be integrated in optical network design to give a computationally tractable metric, which is highly-correlated to maximum achievable throughput. Similar investigations happened around the same time in [20, 164], which focused on the correlation of spectral metrics to the maximum achievable throughput of optical networks, however excluded the traf c distribution within modelling.

With the conclusion of Chapter 4 that DWC is highly correlated to the maximum achievable throughput of networks, Chapter 5 focused on whether optimisation of DWC could directly optimise the maximum achievable throughput of optical networks. The conclusion that it would result in networks that have higher maximum achievable throughput is not an obvious one, since correlations do not necessarily cause changes in the actual target metric. This was investigated by testing several optimisation algorithms, all with the aim of minimising the DWC of networks. A control-set of SNR-BA networks was generated, which generated networks without considering the DWC metric, however was shown in Chapter 3 to give well-performing networks, compared to random networks. The results showed that the networks designed using the DWC objective on average have a 63.1% increase in maximum achievable throughput compared to the control-set. Two other crucial observations were made. Firstly, it was shown that only improving the number of allocatable lightpaths by sacri cing physical properties is not always bene cial, therefore a balance between these two objectives must be found. Secondly, DWC is not a metric for direct throughput optimisation, it promotes a trend of well-performing networks, however amongst the DWC designed networks, the lowest DWC network did not correspond to the best performing network. These results were not only applicable to DWC. Recent work in [164] investigated spectral properties that correlate well with maximum achievable throughput. The authors suggest that following two or more of these metrics can effectively design optical networks for maximum achievable throughput, robustness and cost. However, as seen with DWC, correlational metrics do not directly optimise the maximum achievable throughput. This is because correlations that are measured depend on the data used to calculate them. When designing networks, this data is changed and therefore the correlation measured from previous datasets is not guaranteed within these new networks. The fact that highly correlated design objectives do not result in optimal topologies is important as a wider result, as research focus should be narrowed down to methods that directly optimise structural and physical properties, based on accurate models, as was shown in Chapter 5. The research in this PhD, showed an alternative way to correlational metrics that results in

directly altering the structures and physical properties responsible for maximum achievable throughput. For optical network operators this is important, as implementing optical network designs in reality is immensely costly. Therefore, to maximise the return on investment accurate models and optimisations are required.

Therefore, to achieve direct optimisation of maximum achievable throughput, limiting cut theory combined with GN-modelling and a greedy heuristic was used. This approach showed that the networks designed via limiting cut theory, provide the best maximum achievable throughput with a 106% increase compared to the control-set. However, in some cases resulted in the worst DWC amongst the designed networks. It is concluded that this objective is one of the only methods, other than ILP formulations, to directly optimise optical networks for maximising achievable throughput. This method was then used within a deep reinforcement learning algorithm to demonstrate more advanced optimisation using this objective. It was shown that both through better optimisation and more accurate objective modelling, the maximum achievable throughput can be increased by 148% compared to the control-set and 14% over the greedy-cut method. More importantly, a 27% increase of the worst performing topology compared to the greedy-cut method was observed. This work showed that optimisation of metrics that correlate to the maximum achievable throughput of optical networks, struggle to fully optimise networks. The inclusion of the limiting cut with the GN-model can overcome this and shows an alternate approach to DWC and work done in [20, 164]. In addition, this method has been used within upper-bound estimation for maximum achievable throughput as in [13] and compliments the linear program investigated in section 4.1. The research presented here can be adopted by operators to evaluate future network designs and maximise the achievable throughput of optical networks, whilst also balancing other design objectives, such as latency, resilience and cost.

In overarching conclusion, this thesis provides improved intelligent physical topology design for maximising achievable throughput. Maximising achievable throughput of optical core networks is key to enabling future arti cial intelligence growth. To maximise achievable throughput within the design of optical networks, the role of both structural and physical properties was analysed and included in indirect and direct optimisation objectives. Direct optimisation of maximum achievable throughput was shown through a mixture of greedy optimisation of graph-cuts and the network's physical layer. The intelligence in this solution lies in the intentional optimisation of the structural and physical properties of the network that maximise the achievable throughput of the network, whilst also fully adaptable to varying distance scales and traf c pro les. The successful inclusion of this objective in deep reinforcement learning demonstrates the potential of this objective and its use in more complex optimisation. This work laid the foundations for future network design

through improved direct combinatorial optimisation.

6.1 Future Work

The work presented in this thesis provided major insights into the design of optical networks with the objective of maximising their maximum achievable throughput. This NP-hard optimisation target left many areas of work to still be optimised for the future.

6.1.1 Accurate and Computationally Ef cient Limiting Cut

In Chapter 5, the limiting cut theory within a greedy algorithm was exploited to maximise the maximum achievable throughput of networks. A greedy heuristic was used, due to the computational dif culty of evaluating the limiting cut within more complex combinatorial optimisation frameworks. In addition, networks with 20 nodes were evaluated, however for the future, larger networks need to be designed for datacentres and AI applications. As networks grow ($|N|$), the number of cuts to evaluate over the network grows with $2^{\frac{|N|^2}{2}}$. To use more advanced combinatorial optimisation and to design larger networks more computationally ef cient, accurate and fast methods for evaluating larger limiting cuts of networks need to be investigated. Other authors have started this process already [165, 166]. The accuracy of these methods needs to be validated and then can be integrated into the design process of larger network design and more complex combinatorial optimisation.

6.1.2 Generalising Reinforcement Learning for Physical Topology Design using the Limiting Cut

In section 5.4.4 a deep reinforcement learning agent was trained to maximise achievable throughput within optical network design. This methodology showed 14% average increase over the best performing greedy-cut method, with a 27% increase in the worst performing networks. This demonstrated future potential for using arti cially learnt algorithms for future network design. To make this useful for network operators and to make use of the large amount of training iterations required for this performance, the agent needs to generalise to a variety of scenarios.

Future work would include generalising this agent to different network sizes, different traf c pro les and distance scales. Scaling this design method to large-scale network design (>1000 nodes) would require thorough research in scaling the limiting cut calculation to large networks, as mentioned in section 6.1.1. This research could be used as part of other network designs too, e.g. intra/inter datacentre design.

6.1.3 Learning Generalisable Graph Representations for Optical Networks

One of the methodologies investigated for ef cient maximum achievable throughput calculation was geometric deep learning. A major drawback was that it could not generalise to graph structures different from those trained on. This is a common theme within graph machine learning frameworks, where models are not able to generalise to data drawn from different distributions from those trained upon. An assumption made within the training of the MPNN is that the data used for inference will be in-distribution with respect to the training data. This means that node and edge features, as well as structural and physical properties of these networks are from the same convex hull for training and testing. When this occurs, then the inference is performing interpolation. However, when data is not from the same convex hull as the training data, then extrapolation occurs, which is inference outside of "knowledge space" of the model. Recent research has shown actually that for high-dimension data (100+), the model almost exclusively performs extrapolation [167]. Therefore, to improve the probability that the model operates within the interpolation regime, the dataset needs to grow exponentially in size and diversity with respect to its dimensions. Researching ways to improve the generalisability of these models through training dataset manipulation, training and regularisation strategies is important not only for application to optical networks. These are problems that would be vital to investigate for other application areas even, e.g. neuroscience.

6.1.4 Modelling of Structural versus Physical Properties Performance Trade-off

In Chapter 5, the improvement of maximum achievable throughput within network design was explored in detail. The improvement of maximum achievable throughput originated from both an increase in number of allocated lightpaths (structural) and a decrease in path lengths (physical). Improving the limiting cut of a network, allowed for more lightpaths to be allocated and in turn led to an increase in maximum achievable throughput. However, when improving the limiting cut with detrimental effect on the length of the allocated lightpaths, maximum achievable throughput did not always increase. The exact trade-off between these two properties would characterise for example: how much does the limiting cut need to improve to off-set a certain additional path length. Quantifying this trade-off would help identify, to which extent physical properties can be sacri ced for the reduction of wavelength requirements. Although the effect of structural and physical properties has been studied extensively in this thesis and in [7, 20, 164, 19, 13], this exact trade-off has not been characterised.

6.1.5 Realistic Traf c Modelling

Throughout this thesis the approach to traf c was to use uniform traf c and simple traf c skews. The skews investigated were introduced to give a parametrised skew to the traf c matrix, whilst maintaining total traf c loads. The traf c matrices used, do not represent real-world traf c, as this is dependent on a variety of factors, from population statistics to data-centre locations and size. Additionally, the evolution in network applications such as virtual/augmented reality and arti cial intelligence training requirements are placing signi cant growing loads on core networks and changing the way traf c is distributed across the network. The traf c is a large factor in the design of optical networks and something that affects the maximum achievable throughput. Therefore, realistic traf c modelling, ideally based on real-world data would have to be investigated. A traf c model including data centre locations and traf c was investigated in [168], however this focused on speci c topologies. A model able to handle arbitrary topology and skew is needed. In data-centre networking, a package has been developed for modelling traf c distributions and something similar for optical core networking is essential [169].

6.1.6 Incorporating Modern Technology Constraints

Numerous new technologies are changing the way that optical networks are operated, such as variable bandwidth transceivers, software de ned networking, spatial division multiplexing and Ultra-wideband transmission. These more advanced technologies translate into more dif cult operational optimisation problems, where the routing and wavelength assignment problem is translated into the routing and spectrum assignment problem, which is more dif cult due to the larger number of frequency slots, however also the added spectrum contiguity constraint. Dynamic network scenarios that require routing and spectrum assignment for many requests that arrive and depart in the network on smaller time-scales (minutes/seconds) complicate the operation of designed networks further. Spatial division multiplexing and ultra-wideband transmission expand the number of wavelengths within the networks or add additional constraints to the operation, depending on whether multi-core/mode bres are used. All these maturing technologies need to be taken into account in the network design stage to understand their impact on the network, rather than just in point-to-point transmission. In particular, their impact on the maximum achievable throughput of optical networks needs to be included in future work.

6.1.7 Scalable Integer Linear Programming Design for Maximum Achievable Throughput

In the beginning of Chapter 5, ILP methods for the design of optical networks were discussed. To-date, these are the only algorithms designed to give provably optimal optical networks. Although the worst-case computational complexity of ILP formulations are analogous to that of brute-forcing the solution, however for many instances the optimal solution is found much quicker. These ILPs are focused on cost and minimising wavelength requirements. Combining ILP solutions from sections 5.1 and 4.1.1, could extend an ILP solution that solves the optimal optical network with maximum achievable throughput. In addition, section 4.1.3 investigated linear programs for maximum achievable throughput, which could make the ILP objective more computationally ef cient. An investigation into running these ILPs and improving their ef ciency for larger network sizes is important.

Bibliography

[1] Cisco. Annual Internet Report (2018–2023). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html (2018). Accessed: 12/01/2024.

[2] Ericsson. Mobile data traffic outlook. https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-forecast (2023). Accessed: 18/10/2024.

[3] Ciena. Networks will shape the future of Artificial Intelligence - Ciena. https://www.ciena.com/insights/blog/2024/networks-will-shape-the-future-of-artificial-intelligence (2024). Accessed: 17/10/2024.

[4] ITU. Infrastructure Connectivity Map. https://bbmaps.itu.int/bbmaps/ (2024). Accessed: 19/09/2024.

[5] Puttnam, B. J. et al. 402 Tb/s GMI Data-Rate OESCLU-Band Transmission. In Optical Fiber Communication Conference (OFC) 2024, Th4A.3 (Optica Publishing Group, San Diego California, 2024).

[6] Shannon, C. E. A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379–423 (1948).

[7] Vincent, R. J., Ives, D. J. and Savory, S. J. Scalable Capacity Estimation for Nonlinear Elastic All-Optical Core Networks. Journal of Lightwave Technology 37, 5380–5391 (2019).

[8] Simmons, J. M. Optical Network Design and Planning. Optical Networks (Springer International Publishing, Cham, 2014).

[9] Simon, H. A. and Newell, A. Heuristic Problem Solving: The Next Advance in Operations Research. Operations Research 6, 1–10 (1958).

[10] Tornatore, M., Maier, G. and Pattavina, A. WDM Network Design by ILP Models Based on Flow Aggregation. IEEE/ACM Transactions on Networking 15, 709–720 (2007).

[11] Liu, H. and Tobagi, F. A. Physical Topology Design for All-Optical Networks. In 2006 3rd International Conference on Broadband Communications, Networks and Systems, 1–10 (2006).

[12] Matzner, R., Luo, R., Zervas, G. and Bayvel, P. Intelligent performance inference: A graph neural network approach to modeling maximum achievable throughput in optical networks. APL Machine Learning 1, 026112 (2023).

[13] Cruzado, K., Mori, Y., Lin, S.-C., Matsuura, M., Subramaniam, S. and Hasegawa, H. Capacity-Bound Evaluation and Routing and Spectrum Assignment for Elastic Optical Path Networks with Distance-Adaptive Modulation. In Optical Fiber Communication Conference (OFC) 2024, M3C.6 (Optica Publishing Group, San Diego California, 2024).

[14] Bellman, R. Dynamic programming and stochastic control processes. Information and Control 1, 228–239 (1958).

[15] Sutton, R. S. and Barto, A. G. Reinforcement learning: an introduction. Adaptive computation and machine learning (MIT Press, Cambridge, Mass, 1998).

[16] Mnih, V. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013).

[17] Nevin, J. W., Nallaperuma, S., Shevchenko, N. A., Shabka, Z., Zervas, G. and Savory, S. J. Techniques for applying reinforcement learning to routing and wavelength assignment problems in optical fiber communication networks. Journal of Optical Communications and Networking 14, 733 (2022).

[18] Baroni, S. and Bayvel, P. Wavelength requirements in arbitrarily connected wavelength-routed optical networks. Journal of Lightwave Technology 15, 242–251 (1997).

[19] Cruzado, K., Mori, Y., Lin, S.-C., Matsuura, M., Subramaniam, S. and Hasegawa, H. Effective Capacity Estimation Based on Cut-Set Load Analysis in Optical Path Networks. In 2023 International Conference on Photonics in Switching and Computing (PSC), 1–3 (IEEE, Mantova, Italy, 2023).

[20] Higashimori, K., Inoue, T., Tanaka, T., Inuzuka, F. and Ohara, T. Impact of Physical Topology Features on Performance of Optical Backbone Networks. In 2022 International Conference on Optical Network Design and Modeling (ONDM), 1–6 (IEEE, Warsaw, Poland, 2022).

[21] Higashimori, K., Inuzuka, F. and Ohara, T. Physical topology optimization for highly reliable and ef cient wavelength-assignable optical networks. Journal of Optical Communications and Networking 14, 16 (2022).

[22] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. and Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv:1704.01212 [cs] (2017).

[23] Rusek, K. and Cho da, P. Message-Passing Neural Networks Learn Little's Law. arXiv:1901.05748 [cs] (2019).

[24] Rusek, K., Suarez-Varela, J., Almasan, P., Barlet-Ros, P. and Cabellos-Aparicio, A. RouteNet: Leveraging Graph Neural Networks for Network Modeling and Optimization in SDN. IEEE Journal on Selected Areas in Communications 38, 2260–2270 (2020).

[25] Çetinkaya, E. K., Alenazi, M. J. F., Cheng, Y., Peck, A. M. and Sterbenz, J. P. G. A comparative analysis of geometric graph models for modelling backbone networks. Optical Switching and Networking 14, 95–106 (2014).

[26] Châtelain, B., Bélanger, M. P., Tremblay, C., Gagnon, F. and Plant, D. V. Topological Wavelength Usage Estimation in Transparent Wide Area Networks. IEEE/OSA Journal of Optical Communications and Networking 1, 196–203 (2009).

[27] Agrawal, G. P. Nonlinear Fiber Optics Lecture Notes in Physics (Springer, Berlin, Heidelberg, 2000).

[28] Antonelli, C., Shtaif, M. and Mecozzi, A. Modeling of Nonlinear Propagation in Space-Division Multiplexed Fiber-Optic Transmission. Journal of Lightwave Technology 34, 36–54 (2016).

[29] Semrau, D., Killey, R. I. and Bayvel, P. A Closed-Form Approximation of the Gaussian Noise Model in the Presence of Inter-Channel Stimulated Raman Scattering. Journal of Lightwave Technology 37, 1924–1936 (2019).

[30] Buglia, H., Sillekens, E., Vasylchenkova, A., Bayvel, P. and Galdino, L. On the impact of launch power optimization and transceiver noise on the performance of ultra-wideband transmission systems [Invited]. Journal of Optical Communications and Networking 14, B11 (2022).

[31] Ramaswami, R., Sivarajan, K. N. and Sasaki, G. H. Optical networks: a practical perspective The Morgan Kaufmann series in networking (Elsevier/Morgan Kaufmann, Amsterdam ; Boston, 2010), 3rd ed edn.

[32] Poggiolini, P., Carena, A., Curri, V., Bosco, G. and Forghieri, F. Analytical Modeling of Nonlinear Propagation in Uncompensated Optical Transmission Links. IEEE Photonics Technology Letters 23, 742–744 (2011).

[33] Poggiolini, P. The GN Model of Non-Linear Propagation in Uncompensated Coherent Optical Systems. Journal of Lightwave Technology 30, 3857–3879 (2012). Conference Name: Journal of Lightwave Technology.

[34] Agrawal, G. P. Fiber-optic communication systems No. 222 in Wiley series in microwave and optical engineering (Wiley, New York, 2010), 4th ed edn.

[35] Chlamtac, I., Ganz, A. and Karmi, G. Purely optical networks for terabit communication. In IEEE INFOCOM '89, Proceedings of the Eighth Annual Joint Conference of the IEEE Computer and Communications Societies, 887–896 vol.3 (1989).

[36] Yen, J. Y. Finding the K Shortest Loopless Paths in a Network. Management Science 17, 712–716 (1971).

[37] Napoli, A., Nölle, M., Ra que, D., Fischer, J. K., Spinnler, B., Rahman, T., Mezghanni, M. M. and Bohn, M. On the next generation bandwidth variable transponders for future exible optical systems. In 2014 European Conference on Networks and Communications (EuCNC), 1–5 (IEEE, 2014).

[38] Lord, A., Wright, P. and Mitra, A. Core networks in the exgrid era. Journal of Lightwave Technology 33, 1126–1135 (2015).

[39] Velasco, L., Klinkowski, M., Ruiz, M. and Comellas, J. Modeling the routing and spectrum allocation problem for exgrid optical networks. Photonic Network Communications 24, 177–186 (2012).

[40] Castro, A., Velasco, L., Ruiz, M., Klinkowski, M., Fernández-Palacios, J. P. and Careglio, D. Dynamic routing and spectrum (re) allocation in future exgrid optical networks. Computer Networks 56, 2869–2883 (2012).

[41] Dijkstra, E. A Note on Two Problems in Connexion with Graphs. Numerische Mathematik 269–271 (1959).

[42] Bellman, R. On a Routing Problem. Quarterly of Applied Mathematics Vol. 16 (1958).

[43] Floyd, R. W. Algorithm 97: shortest path. Communications of the ACM 5, 345–345 (1962).

[44] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. Introduction to algorithms (MIT Press, Cambridge, Massachusetts London, England, 2009), third edition edn.

[45] Eppstein, D. Finding the k Shortest Paths. Donald Bren School of Information and Computer Science (1997).

[46] Papadimitriou, C. H. On the complexity of integer programming. Journal of the ACM 28, 765–768 (1981).

[47] Dantzig, G. B. Application of the simplex method to a transportation problem. In Koopmans, T. C. (ed.) Activity Analysis of Production and Allocation, 359–373 (John Wiley and Sons, New York, 1951).

[48] Dantzig, G. B. Linear Programming. Operations Research 50, 42–47 (2002).

[49] Land, A. H. and Doig, A. G. An Automatic Method of Solving Discrete Programming Problems. Econometrica 28, 497 (1960).

[50] Minsky, M. Steps toward Arti cial Intelligence. Proceedings of the IRE 49, 8–30 (1961).

[51] Mokhtar, A. and Azizoglu, M. Adaptive wavelength routing in all-optical networks. IEEE/ACM Transactions on Networking 6, 197–206 (1998).

[52] Ramaswami, R. and Sivarajan, K. Routing and wavelength assignment in all-optical networks. IEEE/ACM Transactions on Networking 3, 489–500 (1995).

[53] Banerjee, D. and Mukherjee, B. A practical approach for routing and wavelength assignment in large wavelength-routed optical networks. IEEE Journal on Selected Areas in Communications 14, 903–908 (1996).

[54] Narula-Tam, A., Lin, P. and Modiano, E. Ef cient routing and wavelength assignment for recon gurable WDM networks. IEEE Journal on Selected Areas in Communications 20, 75–88 (2002).

[55] Bellman, R. and Dreyfus, S. Dynamic Programming, vol. 33 (Princeton University Press, 2010).

[56] Barto, A. G., Sutton, R. S. and Anderson, C. W. Neuronlike adaptive elements that can solve dif cult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics SMC-13, 834–846 (1983).

[57] Watkins, C. J. C. H. and Dayan, P. Q-learning. Machine Learning 8, 279–292 (1992).

[58] Chen, X., Li, B., Proietti, R., Lu, H., Zhu, Z. and Yoo, S. J. B. DeepRMSA: A Deep Reinforcement Learning Framework for Routing, Modulation and Spectrum Assignment in Elastic Optical Networks. Journal of Lightwave Technology 37, 4155–4163 (2019).

[59] Stampa, G., Arias, M., Sánchez-Charles, D., Muntés-Mulero, V. and Cabellos, A. A deep-reinforcement learning approach for software-de ned networking routing optimization. arXiv preprint arXiv:1709.07080 (2017).

[60] Natalino, C., Raza, M. R., Öhlen, P., Batista, P., Santos, M., Wosinska, L. and Monti, P. Machine-learning-based routing of qos-constrained connectivity services in optical transport networks. In Photonic Networks and Devices, NeW3F–5 (Optica Publishing Group, 2018).

[61] Almasan, P., Suárez-Varela, J., Badia-Sampera, A., Rusek, K., Barlet-Ros, P. and Cabellos-Aparicio, A. Deep reinforcement learning meets graph neural networks: An optical network routing use case. arXiv preprint arXiv:1910.07421 (2019).

[62] Suarez-Varela, J., Mestres, A., Yu, J., Kuang, L., Feng, H., Cabellos-Aparicio, A. and Barlet-Ros, P. Routing in optical transport networks with deep reinforcement learning. Journal of Optical Communications and Networking 11, 547–558 (2019).

[63] Shimoda, M. and Tanaka, T. Deep Reinforcement Learning-based Spectrum Assignment with Multi-metric Reward Function and Assignable Boundary Slot Mask. In 2021 Opto-Electronics and Communications Conference (OECC), 1–3 (2021).

[64] Shimoda, M. and Tanaka, T. Mask RSA: End-To-End Reinforcement Learning-based Routing and Spectrum Assignment in Elastic Optical Networks. In 2021 European Conference on Optical Communication (ECOC), 1–4 (2021).

[65] Zhao, Z., Zhao, Y., Ma, H., Li, Y., Rahman, S., Han, D., Zhang, H. and Zhang, J. Cost-ef cient routing, modulation, wavelength and port assignment using reinforcement learning in optical transport networks. Optical Fiber Technology 64, 102571 (2021).

[66] Arce, S., Albertini, L. A., Ríos, I., Pinto-Roa, D. P., Colbes, J. and Villagra, M. Reinforcement Learning applied to the Routing and Spectrum Assignment in Elastic Optical Networks. In 2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 1–6 (2022).

[67] Cicco, N. D., Mercan, E. F., Karandin, O., Ayoub, O., Troia, S., Musumeci, F. and Tornatore, M. On Deep Reinforcement Learning for Static Routing and Wavelength Assignment. IEEE Journal of Selected Topics in Quantum Electronics 28, 1–12 (2022).

[68] Quang, H. T., Houidi, O., Errea-Moreno, J., Verchere, D. and Zeghlache, D. MAGC-RSA: Multi-Agent Graph Convolutional Reinforcement Learning for Distributed Routing and Spectrum Assignment in Elastic Optical Networks. In 2022 European Conference on Optical Communication (ECOC), 1–4 (2022).

[69] Tang, B., Huang, Y.-C., Xue, Y. and Zhou, W. Heuristic Reward Design for Deep Reinforcement Learning-Based Routing, Modulation and Spectrum Assignment of Elastic Optical Networks. IEEE Communications Letters 26, 2675–2679 (2022).

[70] Terki, A. B., Pedro, J., Eira, A., Napoli, A. and Sambo, N. Routing and Spectrum Assignment Assisted by Reinforcement Learning in Multi-band Optical Networks. In European Conference on Optical Communication (ECOC) 2022 (2022), paper Tu5.63, Tu5.63 (Optica Publishing Group, 2022).

[71] Xu, L., Huang, Y.-C., Xue, Y. and Hu, X. Deep Reinforcement Learning-Based Routing and Spectrum Assignment of EONs by Exploiting GCN and RNN for Feature Extraction. Journal of Lightwave Technology 40, 4945–4955 (2022). Conference Name: Journal of Lightwave Technology.

[72] Tanaka, T. and Shimoda, M. Pre- and post-processing techniques for reinforcement-learning-based routing and spectrum assignment in elastic optical networks. Journal of Optical Communications and Networking 15, 1019 (2023).

[73] Frank, H., Frisch, I. T., Van Slyke, R. and Chou, W. S. Optimal design of centralized computer networks. Networks 1, 43–57 (1971).

[74] Gerla, M. and Kleinrock, L. On the Topological Design of Distributed Computer Networks. IEEE Transactions on Communications 25, 48–60 (1977).

[75] Ives, D. J., Bayvel, P. and Savory, S. J. Routing, modulation, spectrum and launch power assignment to maximize the traf c throughput of a nonlinear optical mesh network. Photonic Network Communications 29, 244–256 (2015).

[76] Alvarado, A., Ives, D. J., Savory, S. J. and Bayvel, P. On the Impact of Optimal Modulation and FEC Overhead on Future Optical Networks. Journal of Lightwave Technology 34, 2339–2352 (2016).

[77] Wan, X., Hua, N. and Zheng, X. Dynamic Routing and Spectrum Assignment in Spectrum-Flexible Transparent Optical Networks. Journal of Optical Communications and Networking 4, 603 (2012).

[78] Walkowiak, K., Klinkowski, M. and Lechowicz, P. Dynamic routing in spectrally spatially exible optical networks with back-to-back regeneration. IEEE/OSA Journal of Optical Communications and Networking 10, 523–534 (2018).

[79] Martin, I., Troia, S., Hernandez, J. A., Rodriguez, A., Musumeci, F., Maier, G., Alvizu, R. and Gonzalez de Dios, O. Machine Learning-Based Routing and Wavelength Assignment in Software-De ned Optical Networks. IEEE Transactions on Network and Service Management 16, 871–883 (2019).

[80] Choi, H., Subramaniam, S. and Choi, H.-A. On double-link failure recovery in WDM optical networks. In Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, 808–816 vol.2 (2002).

[81] Moharrami, M., Fallahpour, A., Beyranvand, H. and Salehi, J. A. Resource Allocation and Multicast Routing in Elastic Optical Networks. IEEE Transactions on Communications 65, 2101–2113 (2017).

[82] Archambault, E., Alloune, N., Furdek, M., Xu, Z., Tremblay, C., Muhammad, A., Chen, J., Wosinska, L., Littlewood, P. and Belanger, M. P. Routing and Spectrum Assignment in Elastic Filterless Optical Networks. IEEE/ACM Transactions on Networking 24, 3578–3592 (2016).

[83] Shirin Abkenar, F. and Ghaffarpour Rahbar, A. Study and Analysis of Routing and Spectrum Allocation (RSA) and Routing, Modulation and Spectrum Allocation (RMSA) Algorithms in Elastic Optical Networks (EONs). Optical Switching and Networking 23, 5–39 (2017).

[84] Barabasi, A.-L. and Albert, R. Emergence of scaling in random networks. Science 286, 509–512 (1999).

[85] Fan, C. Spectral Graph Theory (AMS, 2006).

[86] Fay, D., Haddadi, H., Uhlig, S., Moore, A., Mortier, R. and Jamakovic, A. Weighted spectral distribution. IEEE/ACM Transactions on Networking (2008).

[87] Solomonoff, R. and Rapoport, A. Connectivity of random nets. The bulletin of mathematical biophysics 13, 107–117 (1951).

[88] Erdos, P. and Renyi, A. On the Evolution of Random Graphs. Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 17–61 (1960).

[89] Penghui Yuan and Anshi Xu. The In uence of Physical Network Topologies on Wavelength Requirements in Optical Networks. Journal of Lightwave Technology, 28, 1338–1343 (2010).

[90] Waxman, B. Routing of multipoint connections. IEEE Journal on Selected Areas in Communications, 6, 1617–1622 (1988).

[91] Matula, D. W. and Sokal, R. R. Properties of Gabriel Graphs Relevant to Geographic Variation Research and the Clustering of Points in the Plane. Geographical Analysis, 12, 205–222 (1980).

[92] Gabriel, K. R. and Sokal, R. R. A New Statistical Approach to Geographic Variation Analysis. Systematic Zoology, 18, 259 (1969).

[93] Fenger, C., Limal, E., Gliese, U. and Mahon, C. J. Statistical Study of the Correlation Between Topology and Wavelength Usage in Optical Networks with and without Conversion. In Pujolle, G., Perros, H., Fdida, S., Körner, U. and Stavrakakis, I. (eds.) Networking 2000 Broadband Communications, High Performance Networking, and Performance of Communication Networks, Lecture Notes in Computer Science, 168–175 (Springer, Berlin, Heidelberg, 2000).

[94] Wu, H., Zhou, F., Zhu, Z. and Chen, Y. Interference-and-security-aware distance spectrum assignment in Elastic Optical Networks. 2016 21st European Conference on Networks and Optical Communications (NOC), 100–105 (IEEE, Lisbon, Portugal, 2016).

[95] Ashraf, M. W., Idrus, S. M., Iqbal, F. and Butt, R. A. On spatially disjoint lightpaths in optical networks. Photonic Network Communications, 36, 11–25 (2018).

[96] Pages, A., Perello, J., Spadaro, S. and Junyent, G. Strategies for Virtual Optical Network Allocation. IEEE Communications Letters, 16, 268–271 (2012).

[97] Depizzol, D. B., Montalvão, J., Lima, F. D. O., Moreira Paiva, M. H. and Vieira Segatto, M. E. Feature selection for optical network design via a new mutual information estimator. Expert Systems with Applications, 107, 72–88 (2018).

[98] Semrau, D., Durrani, S., Zervas, G., Killey, R. I. and Bayvel, P. On the Relationship Between Network Topology and Throughput in Mesh Optical Networks. arXiv:2008.06708 [cs, eess] (2020). ArXiv: 2008.06708.

[99] Pavan, C., Morais, R. M., Ferreira da Rocha, J. R. and Pinto, A. N. Generating Realistic Optical Transport Network Topologies. IEEE/OSA Journal of Optical Communications and Networking 2, 80–90 (2010).

[100] Velinska, J., Mirchev, M. and Mishkovski, I. Optical networks' topologies: costs, routing and wavelength assignment. Optical networks 1–10 (2017).

[101] Poggiolini, P., Bosco, G., Carena, A., Curri, V., Jiang, Y. and Forghieri, F. The GN-Model of Fiber Non-Linear Propagation and its Applications. Journal of Lightwave Technology 32, 694–721 (2014).

[102] Orlowski, S., Wessaly, R., Pioro, M. and Tomaszewski, A. SNDlib 1.0—Survivable Network Design Library. Networks 55, 276–286 (2010).

[103] Ramamurthy, B., Feng, H., Datta, D., Heritage, J. and Mukherjee, B. Transparent vs. opaque vs. translucent wavelength-routed optical networks. In OFC/IOOC . Technical Digest. Optical Fiber Communication Conference, 1999, and the International Conference on Integrated Optics and Optical Fiber Communication, 59–61 (IEEE, San Diego, CA, USA, 1999).

[104] De Maesschalck, S. Pan-European Optical Transport Networks: An Availability-based Comparison. Photonic Network Communications 5, 203–225 (2002).

[105] De Maesschalck, S. Network Aspects (NA); Availability performance of path elements of international digital paths. Tech. Rep. REN/NA-042140, European Telecommunications Standards Institute, France (1998).

[106] Nelder, J. A. and Mead, R. A simplex method for function minimization. The computer journal 7, 308–313 (1965).

[107] Marsaglia, G., Tsang, W. W. and Wang, J. Evaluating kolmogorov's distribution. Journal of statistical software 8, 1–4 (2003).

[108] Garcia, N. M., Pereira, M., Freire, M. M., Monteiro, P. P. and Lenkiewicz, P. Performance of Optical Burst Switched Networks for Grid Applications. In International Conference on Networking and Services (ICNS'07), 120–120 (IEEE, Athens, Greece, 2007).

[109] Baroni, S., Gibbens, R. J. and Bayvel, P. On the number of wavelengths in arbitrarily-connected wavelength-routed optical networks. Optical Networks and Their Applications, MN2 (OSA, Washington D. C., 1998).

[110] Essiambre, R.-J., Kramer, G., Winzer, P., Foschini, G. and Goebel, B. Capacity Limits of Optical Fiber Networks. Lightwave Technology, Journal of 28, 662–701 (2010).

[111] Mecozzi, A. and Essiambre, R.-J. Nonlinear Shannon Limit in Pseudolinear Coherent Systems. Journal of Lightwave Technology 30, 2011–2024 (2012). Conference Name: Journal of Lightwave Technology.

[112] Matzner, R., Semrau, D., Luo, R., Zervas, G. and Bayvel, P. Making intelligent topology design choices: understanding structural and physical property performance implications in optical networks [Invited]. Journal of Optical Communications and Networking 13, D53–D67 (2021). Publisher: Optica Publishing Group.

[113] Matzner, R., Ahuja, A., Sadeghi, R., Doherty, M., Savory, S. J. and Bayvel, P. Topology Bench: Systematic Graph Based Benchmarking for Core Optical Networks. Journal of Optical Communications and Networking (2024).

[114] Chlamtac, I., Ganz, A. and Karmi, G. Lightpath communications: an approach to high bandwidth optical WAN's. IEEE Transactions on Communications 40, 1171–1182 (1992).

[115] Yufeng Xin, Rouskas, G. and Perros, H. On the physical and logical topology design of large-scale optical networks. Journal of Lightwave Technology 21, 904–915 (2003).

[116] Bannister, J., Fratta, L. and Gerla, M. Topological design of the wavelength-division optical network. In Proceedings. IEEE INFOCOM '90: Ninth Annual Joint Conference of the IEEE Computer and Communications Societies, 1005–1013 (IEEE Comput. Soc. Press, San Francisco, CA, USA, 1990).

[117] Nagatsu, N., Okamoto, S. and Sato, K. Large scale photonic transport network design based on optical paths. In Proceedings of GLOBECOM'96. 1996 IEEE Global Telecommunications Conference, vol. 1, 321–327 (IEEE, London, UK, 1996).

[118] Jyothi, S. A., Singla, A., Godfrey, P. B. and Kolla, A. Measuring and Understanding Throughput of Network Topologies. SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 761–772 (2016).

[119] Namyar, P., Supittayapornpong, S., Zhang, M., Yu, M. and Govindan, R. A throughput-centric view of the performance of datacenter topologies. In

Proceedings of the 2021 ACM SIGCOMM 2021 Conference, 349–369 (ACM, Virtual Event USA, 2021).

[120] Betker, A., Gamrath, I., Kosiankowski, D., Lange, C., Lehmann, H., Pfeuffer, F., Simon, F. and Werner, A. Comprehensive topology and traf c model of a nationwide telecommunication network. Journal of Optical Communications and Networking 6, 1038–1047 (2014). Conference Name: Journal of Optical Communications and Networking.

[121] Di Bucchianico, A. Coef cient of Determination (R2). In Encyclopedia of Statistics in Quality and Reliability (American Cancer Society, 2008).

[122] Ives, D. J., Alvarado, A. and Savory, S. J. Throughput Gains From Adaptive Transceivers in Nonlinear Elastic Optical Networks. Journal of Lightwave Technology 35, 1280–1289 (2017).

[123] de Araújo, D. R. B., Martins-Filho, J. F. and Bastos-Filho, C. J. A. Using Multi-Layer Perceptron and complex network metrics to estimate the performance of optical networks. In 2013 SBMO/IEEE MTT-S International Microwave Optoelectronics Conference (IMOC), 1–5 (2013).

[124] LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[125] Araujo, D. R. B., Bastos-Filho, C. J. A. and Martins-Filho, J. F. Analyzing surrogate models to assess Blocking Probability of optical networks. 2015 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC), 1–5 (IEEE, Porto de Galinhas, Brazil, 2015).

[126] Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M. and Monfardini, G. The Graph Neural Network Model. IEEE Transactions on Neural Networks 20, 61–80 (2009).

[127] Kipf, T. N. and Welling, M. Semi-supervised classi cation with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).

[128] Li, Y., Tarlow, D., Brockschmidt, M. and Zemel, R. Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493 (2015).

[129] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R. P. Convolutional networks on graphs for learning molecular ngerprints. Advances in neural information processing systems 28 (2015).

[130] Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D. et al. Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems 29 (2016).

[131] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P. Molecular graph convolutions: moving beyond fingerprints. Journal of computer-aided molecular design 30, 595–608 (2016).

[132] Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. Nature communications 8, 13890 (2017).

[133] Bruna, J., Zaremba, W., Szlam, A. and LeCun, Y. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013).

[134] Defferrard, M., Bresson, X. and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems 29 (2016).

[135] Wang, C., Yoshikane, N. and Tsuritani, T. Usage of a Graph Neural Network for Large-Scale Network Performance Evaluation. 2021 International Conference on Optical Network Design and Modeling (ONDM), 1–5 (IEEE, Gothenburg, Sweden, 2021).

[136] Defferrard, M., Bresson, X. and Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. Advances in Neural Information Processing Systems, vol. 29 (Curran Associates, Inc., 2016).

[137] Cho, K., Van Merrienboer, B., Bahdanau, D. and Bengio, Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 103–111 (Association for Computational Linguistics, Doha, Qatar, 2014).

[138] Luo, R., Xu, Y.-Z., Matzner, R., Zervas, G., Saad, D. and Bayvel, P. Message passing: towards low-complexity, global optimal routing and wavelength assignment solutions for optical networks. Optical Fiber Communication Conference, Th1F–5 (Optica Publishing Group, 2022).

[139] Bouillet, E., Ellinas, G., Labourdette, J.-F. and Ramamurthy, R. Path Routing in Mesh Optical Networks (John Wiley and Sons, 2007).

[140] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* **32**, 4–24 (2020).

[141] Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).

[142] Estrada, E., Higham, D. and Hatano, N. Communicability Betweenness in Complex Networks. *Physica A: Statistical Mechanics and its Applications* **388** (2009).

[143] Xu, Y.-Z. and Saad, D. Network pruning and growth: Probabilistic optimization. *Physical Review Research* **5**, 033087 (2023).

[144] Labourdette, J.-F., Acampora, A. and Hart, G. Reconfiguration algorithms for rearrangeable lightwave networks. In *[Proceedings] IEEE INFOCOM '92: The Conference on Computer Communications*, 2205–2214 vol.3 (IEEE, Florence, Italy, 1992).

[145] Mukherjee, B., Banerjee, D., Ramamurthy, S. and Mukherjee, A. Some principles for designing a wide-area WDM optical network. *IEEE/ACM Transactions on Networking* **4**, 684–696 (1996).

[146] Xiao, G., Leung, Y.-W. and Hung, K.-W. Two-stage cut saturation algorithm for designing all-optical networks. *IEEE Transactions on Communications* **49**, 1102–1115 (2001).

[147] De Araujo, D. R. B., Martins-Filho, J. F. and Bastos-Filho, C. J. A. New Graph Model to Design Optical Networks. *IEEE Communications Letters* **19**, 2130–2133 (2015).

[148] Ko, K.-T., Tang, K.-S., Chan, C.-Y., Man, K.-F. and Kwong, S. Using genetic algorithms to design mesh networks. *Computer* **30**, 56–61 (1997).

[149] Sayoud, H., Takahashi, K. and Vaillant, B. Designing communication network topologies using steady-state genetic algorithms. *IEEE Communications Letters* **5**, 113–115 (2001).

[150] Altiparmak, F., Dengiz, B. and Smith, A. E. Optimal Design of Reliable Computer Networks: A Comparison of Metaheuristics. *Journal of Heuristics* **9**, 471–487 (2003).

[151] Chaves, D. A. R., Bastos-Filho, C. J. A. and Martins-Filho, J. F. Multiobjective physical topology design of all-optical networks considering QoS and Capex. In

2010 Conference on Optical Fiber Communication (OFC/NFOEC), collocated National Fiber Optic Engineers Conference, 1–3 (2010).

[152] Araújo, D. R. B., Bastos-Filho, C. J. A., Barboza, E. A., Chaves, D. A. R. and Martins-Filho, J. F. An efficient multi-objective evolutionary optimizer to design all-optical networks considering physical impairments and CAPEX. In *2011 11th International Conference on Intelligent Systems Design and Applications*, 76–81 (2011). ISSN: 2164-7143.

[153] de Araújo, D. R. B., Bastos-Filho, C. J. A. and Martins-Filho, J. F. An evolutionary approach with surrogate models and network science concepts to design optical networks. *Engineering Applications of Artificial Intelligence* **43**, 67–80 (2015).

[154] Figueiredo, E. M., Araújo, D. R., Filho, C. J. B. and Ludermir, T. B. Physical Topology Design of Optical Networks Aided by Many-Objective Optimization Algorithms. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, 409–414 (2016). ISSN: null.

[155] Kennedy, J. and Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1942–1948 (IEEE, Perth, WA, Australia, 1995).

[156] Holland, J. H. Genetic algorithms. *Scientific american* **267**, 66–73 (1992).

[157] Kachitvichyanukul, V. Comparison of Three Evolutionary Algorithms: GA, PSO, and DE. *Industrial Engineering and Management Systems* **11**, 215–223 (2012).

[158] Abedifar, V. and Eshghi, M. An optimized design of optical networks using evolutionary algorithms. *Journal of High Speed Networks* **20**, 11–27 (2014).

[159] Luo, R., Matzner, R., Ottino, A., Zervas, G. and Bayvel, P. Exploring the relationship among traffic, topology, and throughput: towards a traffic-optimal optical network topology design. *Journal of Optical Communications and Networking* **15**, B1–B10 (2023). Conference Name: Journal of Optical Communications and Networking.

[160] Darvariu, V.-A., Hailes, S. and Musolesi, M. Goal-directed graph construction using reinforcement learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **477**, 20210168 (2021).

[161] Darvariu, V.-A., Hailes, S. and Musolesi, M. Planning spatial networks with Monte Carlo tree search. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **479**, 20220383 (2023).

[162] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

[163] Aouchiche, M. and Hansen, P. Distance Laplacian eigenvalues and chromatic number in graphs. *Filomat* **31**, 2545–2555 (2017).

[164] Higashimori, K., Tanaka, T., Inuzuka, F., Ohara, T. and Inoue, T. Key physical topology features for optical backbone networks via a multilayer correlation analysis. *Journal of Optical Communications and Networking* **15**, B23 (2023).

[165] Tsukiyama, S., Shirakawa, I., Ozaki, H. and Ariyoshi, H. An Algorithm to Enumerate All Cutsets of a Graph in Linear Time per Cutset. *Journal of the ACM* **27**, 619–632 (1980).

[166] Hayashi, K., Mori, Y. and Hasegawa, H. Efficient Network Capacity Expansion by Differentiated WDM-Density with Bandwidth-Variable Virtual Direct Links. In *2022 IEEE Future Networks World Forum (FNWF)*, 310–313 (IEEE, Montreal, QC, Canada, 2022).

[167] Balestriero, R., Pesenti, J. and LeCun, Y. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485* (2021).

[168] Goścień, R. and Walkowiak, K. Modeling and optimization of data center location and routing and spectrum allocation in survivable elastic optical networks. *Optical Switching and Networking* **23**, 129–143 (2017).

[169] Parsonson, C. W., Benjamin, J. L. and Zervas, G. Traffic generation for benchmarking data centre networks. *Optical Switching and Networking* **46**, 100695 (2022).