The Dissertation Committee for Abishek Sankararaman
certifies that this is the approved version of the following dissertation:

Spatial Stochastic Models for Network Analysis

Committee:

François Baccelli, Supervisor

Gustavo deVeciana

Sanjay Shakkottai

Alexandros Dimakis

Joe Neeman

Spatial Stochastic Models for Network Analysis

by

Abishek Sankararaman,

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2019

Dedicated to Amma and Appa, for all their love, blessings and sacrifices.

Acknowledgements

I take great joy in writing this chapter to thank all the people that have made the last few years working on this thesis very memorable. First and foremost, I pay my thanks and respect to my advisor Prof. François Baccelli, without whose support, guidance, goodwill, trust and affection, this thesis would not have been possible. I am very humbled that he accepted me as his student and taught me, both the joy of mathematics as well as the beauty in engineering applications. I have always been amazed, both by his breadth, depth and clarity of thought and knowledge. I will always cherish our Monday morning meetings, be it the technical discussions on the board or broad advice on how to proceed when stuck in research, the meetings were a source of inspiration. Above all, the very humility and grace with which he leads his life has given me a benchmark to lead my own and hope to reach his standard some day.

I also acknowledge my sincere thanks to Prof. Sanjay Shakkottai. He took me on as a Teaching Assistant and then later on as a research collaborator and guided me as his own student. For that, I shall always be very grateful, for I have learnt a lot about research from him the last few years. His infective enthusiasm for research and new ideas makes it an absolute pleasure to work and interact with him. I am also grateful to the wonderful class on probability

he taught that have inspired me to take on some of the projects in this thesis.

I would like to thank Prof. Gustavo deVeciana, for many hours of patient discussion and advice, be it in listening to my talks before a conference presentation and polishing them, or on interviewing and choosing career options. His class on queueing theory in my first year is very memorable to me, as I would spend hours on homeworks only to go to his office hours and be dazzled by his remarkably simple and elegant solutions. His class was very instrumental in my understanding of queueing and dynamics that aided me greatly in my research.

I would also thank Prof. Joe Neeman for serving on my committee and for a wonderful class on Markov Chains. That class sparked my interest in discrete probability and in community detection, which I later had an opportunity to work on in this thesis. I also thank Prof. Alex Dimakis for many valuable inputs during grad school. His passion for exploring research, flair for giving engaging talks and commitment and work ethic have always been a source of inspiration for me.

I also give special thanks to Prof. Sergey Foss, with whom I have had a wonderful research collaboration. He graciously supported me for a two weeks visit to Heriot-Watt university, where much of the work in Chapter 3 took place. He was a great host and took a lot of time and effort, in both helping me get around Scotland and also teaching me the wisdom of mathematics he had acquired over his life. I feel very fortunate to have been associated so closely with him and look forward to future collaborations. I also acknowledge my

grateful to my in-laws for being supportive through this journey,

Most importantly, I take this opportunity to pay my respect and thanks to my parents and grandparents. They have instilled in me, through the way they led their lives, the virtues of honesty and hard work. I am forever indebted to them for the incredibly amazing upbringing and blessing me with both courage and conviction, to travel half way across the world in pursuit of knowledge.

Spatial Stochastic Models for Network Analysis

Abishek Sankararaman, Ph.D.
The University of Texas at Austin, 2019

Supervisor: François Baccelli

This thesis proposes new stochastic interacting particle models for networks, and studies some fundamental properties of these models. This thesis considers two application areas of networking - engineering design questions in future wireless systems and algorithmic tasks in large scale graph structured data. The key innovation introduced in this thesis is to bring tools and ideas from stochastic geometry to bear on the problems in both these application domains. We identify certain fundamental questions in design and engineering both wireless systems and large scale graph structured data processing systems. Subsequently, we identify novel stochastic geometric models, that captures the fundamental properties of these networks, which forms the first research contribution. We then rigorously study these models, by bringing to bear new tools from stochastic geometry, random graphs, percolation and Markov processes to establish structural results and fundamental phase transitions in these models. Using our developed mathematical methodology, we

then identify design insights and develop algorithms, which we demonstrate are instructive in many practical settings.

In the setting of wireless systems, this thesis studies both ad-hoc and cellular networks. In the ad-hoc network setting, we aim to understand fundamental limits of the simplest possible protocol to access the spectrum, namely a link transmits whenever it has data to send by treating all interference as noise. Surprisingly this basic question itself was not understood, as the system dynamics is coupled spatially due to the interference links cause one another and temporally due to randomness in traffic arrivals. We propose a novel interacting particle model called the spatial birth-death wireless network model to understand the stability properties of the simple spectrum access protocol. Using tools from Palm calculus and fluid limit theory, we establish a tight characterization of when this model is stable. Furthermore, we show that whenever stable, the links in steady-state exhibit a form of clustering. Leveraging these structural results, we propose two mean field heuristics to obtain formulas for key performance metrics such as average delay experienced by a link. We empirically find that the proposed formulas for delay predicts accurately the system behavior. We subsequently study scalability properties of this model by introducing an appropriate infinite dimensional version of the model we call the Interference Queueing Networks model. The model consists of a queue located at each grid point of an infinite regular integer lattice, with the queues interacting with each other in a translation invariant fashion. We then prove several structural properties of the model namely, tight conditions

for existence of stationary solutions and some sufficient conditions for uniqueness of stationary solutions. Remarkably, we obtain exact formula for mean delay in this model, unlike the continuum model where we relied on mean-field type heuristics to obtain insights. In the setting of cellular networks, we study optimal association schemes by mobile phones in the case when there are several possible base station technologies operating on orthogonal bands. We show that this choice leads to a performance gain we term technology diversity. Interestingly, we show that the performance gain relies on the amount of instantaneous information a user has on the various base station technologies that it can leverage to make the association decision. We outline optimal association schemes under various information settings that a user may have on the network. Moreover, we propose simple heuristics for association that relies on a user obtaining minimal instantaneous information and are thus practical to implement. We prove that in certain natural asymptotic regime of parameters, our proposed heuristic policy is also optimal, and thus quantifying the value of having fine grained information at a user for association. We empirically observe that the asymptotic result is valid even at finite parameter regimes that are typical in todays networks.

In the application of analyzing large scale graph structured data, we consider the graph clustering problem with side information. Graph clustering is a standard and widely used task which consists in partitioning the set of nodes of a graph into underlying clusters where nodes in the same cluster are similar to each other and nodes across different clusters are different. Moti-

vated by applications in social and biological networks, we consider the task of clustering nodes of a graph, when there is side information on the nodes, other than that contained in the graph. For instance in social networks, one has access to meta data about a person (node in a social graph) such as age, location, income etc, along with the combinatorial data of who are his friends on the social graph. Similarly, in biological networks, there is often meta-data about an experiment that provides additional contextual data about a node, in addition to the combinatorial data. In this thesis, we propose a generative model for such graph structured data with side information, which is inspired by random graph models in stochastic geometry such as the random connection model and the generative models for networks with clusters without contexts, such as the stochastic block model or the planted partition model. We propose a novel graph model called the planted partition random connection model. Roughly speaking, in this model, each node has two labels - an observable $\mathbb{R}^d$ valued (for some fixed d) feature label and an unobservable binary valued community label. Conditional on the node labels, edges are drawn at random in this graph depending on both the feature and community labels of the two end points. The clustering task consists in recovering the underlying partition of nodes corresponding to the respective community labels better than a random assignment, when given an observation of the graph generated and the features of all nodes. We show that if the 'density of nodes', i.e., average number of nodes having features in an unit volume of space of $\mathbb{R}^d$ is small, then no algorithm can cluster the graph that can asymptotically beat a random

assignment of community labels. On the contrary, if the density of nodes is sufficiently high, we give a simple algorithm that recovers the true underlying partition strictly better a random assignment. We then apply the proposed algorithm to a problem in computational biology called Haplotype Phasing and observe empirically, that it obtains state of art results. This demonstrates, both the validity of our generative model, as well as our new algorithm.

Table of Contents

List of Tables

List of Figures

xxii

Chapter 1

Introduction

This thesis proposes and studies, new mathematical models of large networks inspired from recent advances and trends in information technology. We consider two emerging technologies of interest - (i) large scale ad-hoc wireless networks and (ii) graph structured network data with side information. Both these areas are receiving increased amount of attention, both in academic literature and in industrial research, due to emergent technologies such as the drive toward 5G networks in the case of wireless systems and the explosion of available graph structured data from various sources such as social and biological networks, in the case of data analysis. In the topic of 5G networks, this thesis focuses on two key questions. One is the performance analysis of a *simple and scalable protocol* for large scale ad-hoc networks and the other is a problem of user-association in modern cellular networks with several overlapping technologies (such as LTE, 3G and/or several different operators). In the case of large scale graph structured data, the main application of interest in this thesis is the question of community detection, namely identifying underlying latent clusters of groups in large graph structured datasets. We then study a particular case of connection between general community detection and haplotype-phasing, a fundamental problem in computational genomics

that lends itself to the community detection framework we introduce in this thesis.

Wireless networks are getting increasingly complex to engineer, especially with the growing demands and performance requirements placed on them. The advent of new technologies and solution paradigms to meet some of these demands are collectively labeled as *5G technology*. Many challenging research questions then emerges from this drive towards newer technologies and increased performance requirements. A key common feature of modern wireless technologies are that they are large-scale, i.e., a large number of wireless nodes will be in use simultaneously. For example, one can envision that in the near future, a multitude of devices sharing the wireless spectrum such as sensors in the environment and automobiles, wearable devices on people and other monitoring devices on infrastructure, all working together to deliver newer wireless services. The most widely used design abstraction to engineering such systems is to view them as *ad-hoc* wireless networks, i.e., a collection of individual wireless nodes accessing the shared spectrum without any centralized entity controlling them. In this thesis, we ask the question as to how well such ad-hoc networks perform when every node employs the simplest protocol to access the spectrum, namely they transmit whenever they have data to send. Surprisingly, this basic question itself was not fully understood and we make progress towards this in the present thesis. The key challenge in this study arises as wireless is a broadcast medium, and thus the instantaneous rate of service at any node depends on the locations of other concurrently

active nodes, as the other active nodes cause interference. In this thesis, we introduce a novel mathematical model using stochastic geometry and information theory, to capture the essential features of such large scale ad-hoc networks, and in particular capture precisely the dynamics of the network and its impact due to interference. We first introduce this model on a compact Euclidean space which we call the Spatial Birth-Death Wireless network model. Roughly speaking, this is a particle process, where we assume wireless links, each of which is a transmitter and receiver pair, are 'born' in space and 'die' after a while. In other words, links arrive randomly, stay for a duration that is governed by the geometry of other concurrently active links and then exit the network and 'die' after the transmitter completes a certain file transfer. This model naturally couples the temporal evolution of links with the interference experienced by modeling the rate of communication between any transmitter and receiver using results from information theory, when all interference is treated as noise [130].

Subsequently, we extend the above particle dynamics to a certain infinite network setting. The analysis of the spatial birth-death model crucially relies on the fact that the set of all potential locations for a link form a compact subset of the Euclidean plane. This compactness allowed us to use the theory of finite dimensional Markov processes to reason about the dynamics of the spatial birth-death model. However, the results and the analysis from this case do not extend to the limiting case if the set of possible locations is a non compact subset or the entire Euclidean plane. The study of infinite

networks in this context encompasses two goals - (i) to better understand the behavior of 'large' networks, for which an infinite model provides a natural abstraction which is somewhat easier to mathematically analyze on account of the translation invariance of the plane and - (ii) the infinite network sheds light on *scalability* of the proposed protocol, namely the one where every link transmits whenever it has a packet to do so by treating all interference as noise. In this context a protocol is scalable if the outcome for a typical link, such as the average delay experienced by a typical link, remains finite as the network 'size goes to infinity'. Scalability as stated here is indeed a very desirable property, as it implies that a typical user will not be adversely affected by links accessing the spectrum 'far away' in space. However, results from the spatial birth-death model on compact domains is not sufficient to answer this question of scalability and hence we treat the infinite model separately from the spatial birth-death model on compact domains. We introduce a discretized version of the spatial birth-death model on the infinite lattice grid and call this model, the 'Interference Queuing Network' model. We will show in the sequel that such a discretization of the continuum can be done without loss of generality and one can recover the some of the results in the continuum by appropriate limiting procedures. Nevertheless, forming a full analog of results obtained in this discrete infinite model to the continuum infinite model is left as future work. In addition to the discretization, we make an approximation in the way interference impacts the temporal evolution of links. As opposed to the spatial birth-death model, which uses an exact information theoretic formulation for

the rate of file transfer, in the infinite model, we make a low SINR (Signal to Interference plus Noise) approximation, or a linearization of the capacity formula given by information theory. This approximation is made for mathematical tractability and we demonstrate that this nevertheless captures the key features of the model and is capable of delivering practical insights. We conduct a detailed mathematical study to obtain both qualitative and quantitative results on the behavior of the network. We envision that such analytical studies based on the spatial birth death model or the interference queueing network model, can serve as a benchmark to design and evaluate the performance of more complex protocols, such as those involving admission control and/or interference cancellation or other sophisticated network information theoretic schemes.

Similar to the ad-hoc networks case, there is expected to be newer engineering challenges in the cellular networks of the future. In this thesis, we consider the problem of user association in the presence of multiple *orthogonal* cellular services servicing a single user. A concrete example of such a model is the current Google Fi service, where multiple different service providers, operating on orthogonal frequency bands 'pool' together to provide a common service. We consider the problem of how must a mobile user leverage instantaneous information it has about the network, to derive benefits from this 'diversity' in available technologies. Following the convention in information theory, we call the potential gains in performance experienced by a mobile user in such a setting as *technology diversity*. In this case too, we establish

several qualitative results that characterize optimal association schemes that yield the benefits from technology diversity, as well as propose novel heuristics which are near optimal under certain limiting regimes.

In the topic of large graph analysis, the main question studied in this thesis is that of community detection. Roughly speaking, community detection also known as graph clustering consists of partitioning the nodes of a graph into many clusters or classes, such that nodes of the same class are 'similar' to each other and nodes across clusters or communities are 'different'. This is a fundamental problem studied widely in machine learning, statistics and probability literature. This problem has received tremendous attention, both due to the mathematical elegance, and the numerous applications such a clustering paradigm has found. We refer the reader to Chapter 5 for references and the various applications of the community detection paradigm. In this thesis, we consider the question of how well can one perform graph clustering, when the available information is not only the graph, but also additional 'covariate' or side-information about the nodes. Concretely, we study the case when each node as two 'features', an unobserved community label and an observed covariate information, which we model as a vector in some d dimensional Euclidean space. Then the observed graph is a noisy interaction among nodes that depend both on their latent community label and the Euclidean covariate or side-information vector. We will demonstrate in the sequel that such models better capture many applications such as social networks and textual networks of blogs on the web compared to the classical approaches such as

the Stochastic Block Models. In the case of social networks, the side information can correspond to observable features of individuals such as location, age, income etc. and in a network of blogs on the web, the side information can correspond to certain embedding of the text content such as a bag-of-words representation of the text data. In Chapter 5, we introduce a new mathematical model of spatial random graphs with planted communities and consider the community detection problem on this generative model. We establish both qualitative and quantitative results for this model. We identify certain fundamental phase transitions on the model parameters, that separate the regimes from when there exists an algorithm to cluster the graph from regimes when clustering is impossible. We then propose a novel and simple algorithm that can cluster the spatial random graph with communities under further restrictive conditions for clustering than predicted by the phase transition result. Subsequently in Chapter 6, we consider the haplotype phasing problem, a key question in computational genomics ([125],[245]) and demonstrate that this problem can be recast as a community detection question on a spatial random graph. We then empirically establish that the performance of our spatial community detection algorithm is superior to the prior state of art method for haplotype phasing. This demonstrates, both the applicability of the spatial random graph model with planted communities, and the effectiveness of our proposed algorithm compared to state of art methods for clustering that do not exploit such a spatial embedding information.

From a modeling stand point, the key object in this thesis is the sta-

tionary Poisson Point Process (PPP) in an Euclidean space. Although this is a simple and classical object, it nevertheless provides a rich mathematical framework, capable of modeling and studying many different problems and applications, leading to many new insights. The field of stochastic geometry was originally motivated by applications in material sciences, astronomy and biology. As the field of stochastic geometry has matured, it has also found applications lately in image analysis [145] and the design of communication networks. In the case of wireless networks for example, it is now standard to model the locations of wireless links in space using stationary point processes (see for example the books of [42], [187]). In this thesis, we contribute to this framework by proposing a new *dynamic* model of wireless links in Chapter 2 that extends the PPP modeling in space of wireless links to a PPP model of links in 'space-time'. This is the first model that combines the ideas developed in queueing theory along with the ideas in stochastic geometry to propose a new model for wireless networks. We then extend the model in Chapter 3 to consider translation invariant infinite queueing network model, which is conceptually new, both from a mathematical view point and a modeling point of view. Subsequently, in Chapter 4, we use the PPP framework to model the locations of base stations in a cellular network, which is now a widely accepted model for cellular systems ([187]). With regard to data networks, we introduce a new model of spatial random graph with planted communities in Chapter 5, which can be viewed as a multi-type version of the classical random connection model ([303]) of a spatial graph studied widely in stochastic geometry.

This random graph model is new and forms a contribution of this thesis. We show the effectiveness of this model on the task of haplotype phasing from computational biology in Chapter 6, which is an example of a spatial network data set.

From an analytical perspective, this thesis makes advances by bringing to bear new mathematical arguments and tools to the proposed models. In the case of spatial birth-death process, our key analytical contribution is to apply the rate conservation principle from Palm calculus ([49]) and classical fluid limit theory of queueing systems ([135]) to study stability and clustering of the spatial birth-death model. The extension of this model to the case of infinite network in Chapter 3 is also conceptually new and requires the development of new analysis and tools. We use Loyne's type coupling from the past arguments ([251]) along with translation invariance and the mass transport principle ([255]) to reason about the dynamics on infinite space. In Chapter 3, we also establish several new qualitative phenomena on infinite network dynamics using ideas that appeared in the classical interacting particle literature such as the Ising model and Gibbs measures. In the study of community detection, we make novel connections between the statistical task of clustering the graph and structural percolation properties of appropriately constructed spatial random graphs. The arguments in that chapter are new and different from those that are used in the study of stochastic block model (SBM) [16], which is the most popular generative model for community detection.

In summary, the contributions of this thesis to the study of wireless net-

works and graph clustering can be viewed as two fold. First, we propose novel mathematical models to the applications in consideration and argue that they better capture reality of the applications compared to the traditional more widely studied models. Subsequently, the thesis introduces new mathematical tools and algorithms to analyze our proposed model, which forms the technical contribution of this thesis. Thirdly, we validate empirically and through simulations, the benefits and insights obtained by our algorithms and analysis to the applications in consideration, lending validity to our proposed models.

1.0.1 Organization of the Thesis

The first three chapters of the thesis discus problems stemming from wireless networks and the last two discuss the problem of community detection. In Chapter 2, we introduce the Spatial Birth-Death model for an ad-hoc wireless network on a compact domain and study its mathematical properties. This chapter contains both qualitative results such as the stability region and clustering properties of this model and quantitative results such as heuristic formulas for mean delay and numerical simulations to better understand the model. Subsequently in Chapter 3, we consider an appropriate infinite network extension of the Spatial Birth-Death model, which we call the Interference Queueing Network model. Chapter 4 considers the problem of user-association in large cellular networks with technology diversity.

In the next two chapters, we focus on the problem of community detection. In chapter 5, we introduce the new generative model of a spatial ran-

dom graph with planted communities and consider the community detection problem on it. We establish, both phase transition results and also present the new spatial community detection algorithm in this Chapter. Finally in Chapter 6, we consider an application of the community detection problem to haplotype phasing, a key question in the study of computational genomics. We demonstrate empirically, that our Euclidean community detection algorithm performs better than the prior state of art method for haplotype phasing, thereby establishing the effectiveness of our clustering algorithm.

The chapters are organized so that they are self contained and it is possible to read a chapter without referencing other chapters. Nevertheless, Chapters 2 and 3 follow sequentially as they both consider complementary aspects of the same underlying question. Similarly, Chapters 5 and 6 follow sequentially as they both again discuss complementary aspects of the same chapter, namely Chapter 5 introduces the mathematical aspects of the community detection algorithm while Chapter 6 studies the empirical performance of our method. In all chapters, certain proofs are deferred to the Appendix for clarity in exposition.

1.1 Contributions

The various chapters have either been published, accepted or are under preparation to be submitted. Chapter 2 appears as part of the paper [336]. Chapter 3 is accepted to appear and is part of the paper [338]. Chapter 4 is part of paper [340]. Chapter 5 is based on papers [17] and [337].

Chapter 2

Spatial Birth-Death Wireless Networks

2.1 Introduction

In this chapter, we consider the problem of studying the spatial dynamics of Device-to-Device (D2D) or ad-hoc wireless networks. Such wireless networks have received a tremendous amount of attention, due on the one hand to their increasing ubiquity in modern technology and on the other hand to the mathematical challenges in their modeling and performance assessment. Wireless is a broadcast medium and hence the nodes sharing a common spectrum in space interact through the interference they cause to one another. Understanding the limitations due to interference and theoretically optimal protocols in such a static spatial setting has long been considered in network information theory under the interference channel [350]. The full characterization of the interference channel is however a long standing open-problem in network information theory.

In recent years, Stochastic Geometry ([187], [42]) has emerged as a way of assessing performance of wireless links in large-scale networks interacting

---

Parts of this chapter is published in [336]. The author was part of formulating, executing and writing up the results in that paper.

through interference in space. These tools have been very popular to model and analyze wireless system performance for a variety of network architectures including D2D networks, mobile-ad hoc networks [46] and cellular networks [33]. However, the main drawback in these models is that they do not have a notion of temporal interaction and do not allow one to represent random traffic (they usually rely on a "full-buffer" assumption, i.e., every link always has a packet to transmit).

This additional dimension of interaction among wireless links sharing a common spectrum adds to the complexity of their performance analysis but nonetheless is very crucial to understand network performance. Most prior work aiming at studying the temporal interaction of links model spatial interactions through binary on-off behavior encoded by interference or conflict graphs. The temporal interactions are then modeled using queuing theoretic ideas of flow based models (for ex: [85], [243], [276]). Such flow models have a long history in applied mathematics and engineering. They were initially proposed to study dynamic resource allocation in wired networks ([209], [263]), and were subsequently used to model and study wireless networks. Flow based queuing models have inspired many seminal results in applied probability and networks in the past. The main drawback in employing such models in a wireless scenario however is that the spatial and information-theoretic interactions are overly simplified and not captured precisely.

Motivated by this, we propose a new spatial flow model, which uses the continuum space to model link interaction through interference as prescribed

by the information-theoretic setting, and also takes into account the interaction of links across time due to traffic variations. Roughly speaking, our model consists of an interacting particle system in space, where links which is a transmitter-receiver pair arrive in space according to a Poisson Point Process in space-time. The transmitter of each link has a file which it wants to transmit to its corresponding receiver. A link exits the network upon completion of this file transfer. The instantaneous rate at which a transmitter can transmit a file to its receiver is given by the instantaneous Shannon rate, which in turn depends on the geometry of the other transmitters in the network transmitting at that instant to their respective receivers. We study this space-time dynamics to identify a phase-transition in the arrival rate such that each link can be guaranteed to exit in finite time almost surely. The model and the question of phase-transition is formalized in Section 2.2.

The mathematical framework we follow for spatial birth-death processes has been studied in different contexts in the probability literature starting from the work of Preston [308]. In recent years, [172] and [304] have also studied in great detail, the problem of general spatial birth and death process which is the basis of our modeling. From a methodological point of view, the work of [57] is the closest in spirit to our work as it also studies a space-time interacting particle process (of a wireline peer-to-peer network). There are several fundamental differences between the model of [57], which is intrinsically stable, and exhibits repulsion, and our model, which is potentially unstable and which exhibits attraction (clustering). Another difference from [57] is

that the death-rate (defined later) is a linear-function of the state whereas our model is non-linear because of the information-theoretic formulation, thereby making the analysis more challenging. Nevertheless, we use some of the ideas developed in that paper.

From an information-theoretic viewpoint, one can interpret our model and the phase-transition result as a form of dynamic network capacity. Our network model can be interpreted as consisting of arrivals of a single antenna Gaussian additive noise point-to-point channels in space. At each instant of time, the network is a random realization of an interference network operating under the scheme of treating interference as noise. The point-to-point channels exit the network upon completion of a file transfer i.e., with the departures happening in a space-time correlated way determined by our dynamics which in turn is derived from the capacity region of an interference channel under treating interference as noise. The phase-transition results in Theorems 1 and 3 give the maximum rate of arrival that can be supported in the network under the scheme of treating interference as noise. Our model and the framework could potentially be generalized to consider the dynamic capacity of other channels like the Multiple Access Channels or Broadcast channels instead of the Gaussian point-to-point channel considered in this thesis. In these models, each arrival could consist of a single transmitter and multiple receivers or multiple transmitters and a single receiver which form a basic unit of the network. This network can then be modeled to evolve in time through dynamics similar in spirit to Equation (2.4). It is beyond the scope of the thesis however to pose

the problem precisely in the case of multiple access or broadcast channels to derive a phase-transition for dynamic capacity and leave this to future work.

Our model also presents a new form of single server queuing network. Based on our model description in Section 2.2, one can come up with two natural queuing model bounds to study the performance of our model. One can construct a 'worse' system by assuming that there is no distance dependent attenuation and all transmitters contribute the same interference to any receiver. This system will predict larger delays than our original system since the interference is higher. Moreover, since there is no geometry, this upper bound system is equivalent to an $M/M/1$ generalized processor sharing system. On the other hand, to come up with lower bounds for delay, one can totally neglect interference and assume that the different links do not interact at all. This assumption will render our model equivalent to an $M/M/\infty$ system. One of our key insight is that simplifying our model to any of the above two dynamics which neglects spatial structure to provide bounds on delay leads to estimates for delay which are very poor (as demonstrated in Section $V.E$). Thus, we really need to consider the spatial structure as done in Section IV to come up with estimates for delay and performance. The evolution of our model thus presents a novel behavior of stochastic dynamics that cannot be captured by a queuing model that neglects spatial interactions.

From an engineering viewpoint, this work is motivated by emerging interest in applications like Device-to-Device (D2D) networks and Internet of Things (IoT). These two applications can be viewed as an instance of our ab-

stract mathematical model which is more general. D2D is being considered as a viable networking architecture in future cellular standards to improve system capacity by offloading some traffic from base-station to other mobile devices that have the same content. Some of the more important use cases for such offloading are in a crowded setting (like a stadium or a concert) where there is a huge density of mobile devices. Another important application of D2D is in enabling cellular operators to provide "proximity based services". In such settings, a mobile may access content (which we model as files) from nearby mobile users possessing the content (which may be likely owing to geographical and temporal proximity) rather than from a base-station. Such networking architectures are being envisioned to both reduce the load on the base-stations and also to develop new markets for mobile services. Thus, a snapshot of a D2D network will resemble our model with some mobile devices connecting to and downloading files from other mobile devices that are nearby. IoT is another technology gaining momentum due to the vast market opportunities to develop user applications that leverage the IoT network (for instance in tracking sensors for health, security etc). This network also resembles a wireless ad-hoc network with different things communicating occasionally data to each other or to a central access point using the shared wireless medium.

2.1.1   Main Contributions

The main contributions of the present chapter are two-fold -

1. *Stochastic Space-Time Dynamic Model*:

In Section 2.2, we define precisely the mathematical model of the network along with the assumptions we are imposing for the mathematical analysis. This model is one of the contributions of the present chapter as it captures precisely the stochastic interactions and dynamics *both* in space and time. In Section 2.3 we state the main mathematical results of this chapter. In subsection 5.3 we give an *exact* characterization of the time-ergodicity criterion i.e., give an explicit and simple formula to determine the phase-transition for dynamic stability. This notion of stability will be made precise in the sequel in Section 2.2.5 . In section 2.3.2, we prove the intuitive result that, when it exists, the steady-state point process in our model exhibits a form of statistical clustering (made precise later), which is detrimental to performance as it creates higher interference powers at typical receivers than in a network with complete independence. We provide the proof of the ergodicity criterion in the Appendix (Section A.1) which requires the use of point-process theory and in particular Palm calculus and stochastic coupling arguments. Our proof techniques for handling dynamic point-processes are to the best of our knowledge new and potentially useful for analyzing other similar dynamic models of wireless networks. More generally, from an information theoretic perspective, we exhibit a form of dynamic network capacity when treating interference as noise. Our modeling framework could potentially be extended to consider the dynamic network capacity of other more complex channels and pose interesting questions. From a queuing

perspective, we exhibit through our model, a new form spatial queuing which cannot be reduced to any traditional non-spatial queuing network.

2. _Formulas for Delay and System Design Insights_:

We provide an explicit closed form formula to compute the phase-transition for dynamic stability in Section 2.3. The phase-transition result however only provides whether the delay experienced by a typical link is finite or not. In Section 2.4, we propose two formulas to approximately compute the mean-number of links per-unit space and the average delay of a typical link. The simplest heuristic is a first order Poisson approximation which relies on a single intensity parameter and hence cannot take clustering into account. We also propose another heuristic, which is a second order cavity type approximation of the second moment measure [136] of the steady-state point process. We find through simulations, that this heuristic works very well in all regimes. This heuristic is potentially useful to derive explicit approximate formulas for mean delay in other spatio-temporal models. From a practical networking perspective, closed form expressions for delay based on system parameters is very crucial. The formulas for delay provide insight into how to dimension D2D networks in terms of maximum allowable space-time traffic intensity or minimum spectral bandwidth needed to provide mean-delay based guarantees to the links in the ad-hoc network.

2.2 System Model - Birth-Death Process for Wireless Flows

In this section, we describe the mathematical model of the dynamic wireless network which we later analyze. Roughly speaking, our model of a network is one wherein links which are transmitter-receiver pairs arrive into the network which is Euclidean space. Each transmitter of a link has a file it wants to send to its receiver. The speed or rate at which a transmitter can send its file to the receiver is a function of the positions of other transmitters transmitting files to their respective receivers. Upon completion of file transfer, a link departs from the network. We make the above dynamic description of the network more precise in the sequel. In subsection 2.2.1, we describe the continuum network topology. In subsection 2.2.2, we describe the process of link and traffic arrivals into the network. Subsection 2.2.3 gives the precise description of how the instantaneous speed or instantaneous rate of file transfer of a link is affected by the presence of other transmitting links. Finally, in subsection 2.2.4, we put together the preceding parts by compactly describing the arrival-departure dynamics of the wireless links.

2.2.1 Spatial Domain

The wireless links considered in this setup are transmitter-receiver pairs. The network at any point of time consists of a certain number of transmitters each transmitting to its own unique intended receiver. This is also commonly referred to as the "dipole-model" of a D2D ad-hoc wireless network.

The wireless links live in $\mathbf{S} \subset \mathbb{R}^2 = [-Q, Q] \times [-Q, Q]$, a square region of the Euclidean plane where Q is a large but fixed *finite* constant. To avoid edge effects, we identify the opposite edges of the square and wrap it around to form a torus. We denote by $|\mathbf{S}|$ as the area of the region $\mathbf{S}$ which is $4Q^2$. We present the mathematical analysis assuming $\mathbf{S}$ is a square torus as it makes exposition of proof ideas easier.

2.2.2 Links and Traffic Arrival Process

The links arrive into the network as a stationary marked space-time process on $\mathbf{S} \times \mathbb{R}$ with intensity λ. This marked point-process on $\mathbf{S} \times \mathbb{R}$ is denoted by $\mathcal{A}$. An atom $p \in \mathbb{Z}$ of $\mathcal{A}$ represents the receiver and is denoted by (x_p, b_p). $x_p \in \mathbf{S}$ denotes the spatial location of receiver p and $b_p \in \mathbb{R}$ denotes the time of arrival into the network of receiver p. Hence, one can represent the point process $\mathcal{A}$ as $\mathcal{A} = \sum_{p \in \mathbb{Z}} \delta_{(x_p, b_p)}$, where $\delta_{(x,b)}$ refers to the Dirac-mass at $(x, b) \in \mathbf{S} \times \mathbb{R}$. To each point p of $\mathcal{A}$, we associate a vector mark of (y_p, L_p), where $y_p \in \mathbf{S}$ and $L_p \in \mathbb{R}^+$, where y_p refers to the location of the transmitter of receiver p and L_p denotes the file-size which the transmitter of p wants to send to the receiver of p. We refer to the pair $(x_p; y_p)$ as link p whose receiver is in location x_p and transmitter in location y_p. The length of link p is denoted by $T_p := ||x_p - y_p||$.

The set of links present or *alive* in the network at time t is denoted by ϕ_t i.e., $\phi_t = \{(x_1; y_1), ..., (x_{N_t}; y_{N_t})\}$, where N_t is the number of links alive in the network at time t. The exact dynamics describing which links are

present at a particular time t will be specified in the sequel. More formally, $\phi_t = \sum_{i=1}^{N_t} \delta_{x_t}$ is a point-process on $\mathbf{S}$ of receivers marked with the location of their transmitters. We use the terminology "*configuration of links*" to refer to a marked point-process on $\mathbf{S}$ (atoms representing the receiver locations) with its marks (representing its corresponding transmitter locations) in $\mathbf{S}$. We denote by $\phi_t^{Tx} = \{y_1, \cdots, y_{N_t}\}$, the point-process of transmitters present at time t in the network and by $\phi_t^{Rx} = \{x_1, \cdots, x_{N_t}\}$, the point process of receivers at time t in the network.

This arrival process can be seen as an incarnation of links initiating communication in a very dense IoT or a D2D network for instance. When a link has a file to transmit (which comes rarely and randomly in time), a node "switches on" and initiates contact with its receiver. Since the network is dense and arrivals are rare, the spatial locations of links initiating connection can be seen as coming from a space-time point-process which we model as the link arrival process.

2.2.3   Data Rate

The transmitter of each link p has a file of size L_p measured in bits which needs to be communicated to its receiver. The transmitter sends this file to its receiver at a time varying rate given by the instantaneous Shannon rate. Denote by $l(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$, a distance dependent 'path-loss' function which encodes how signal power attenuates with distance. More precisely, $l(r)$ is the received power at distance r from a transmitter transmitting at unit-

power. We can thus, define the rate of file transmission by a transmitter to its receiver as

$$R(x, \phi) = C \log_2 \left(1 + \frac{l(||x - y||)}{\mathcal{N}_0 + \sum_{u \in \phi^{Tx} \backslash \{y\}} l(||x - u||)} \right). \qquad (2.1)$$

In the above expression, C is a constant with units in bits per unit time, $\mathcal{N}_0$ denotes the thermal noise power at the receiver, $\sum_{u \in \phi^{Tx} \backslash \{y\}} l(||x - u||)$ denotes the interference seen at location x due to configuration ϕ and $l(||x - y||)$ is the received signal power at x from y. The interference at location x is the sum of attenuated powers from the transmitters in $\phi^{Tx} \backslash \{y\}$ which is the sum of attenuated powers from all other transmitters other than the transmitter of the tagged receiver under consideration. For any $(x; y) \in \phi$, denote by $I(x, \phi)$ as the interference seen at x in configuration ϕ, which can be written as

$$I(x, \phi) = \sum_{u \in \phi^{Tx} \backslash \{y\}} l(||x - u||). \qquad (2.2)$$

Further, denote by a the constant (which can possibly be infinite) $a = \int_{x \in \mathbf{S}} l(||x||) dx$.

Some common examples of path-loss functions are

- $l(r) = r^{-\alpha}$ with $\alpha > 2$ called the "power-law path-loss" model.

- $l(r) = (r + k)^{-\alpha}$ where k is a constant is commonly called the "bounded path-loss" model.

In our analysis however, we remain general and do not explicitly assume a particular form for the function $l(\cdot)$. Equation (2.1) is the Shannon formula

for the Gaussian SISO (Single Input Single Output) channel with signal power 1 and the interference treated as noise [130].

In Equation (2.1), we did not consider the effect of random channel fading. However, one can easily model the effect of fast fading by defining the rate-function as

$$R^{(f)}(x, \phi) = C\mathbb{E}_h \left[\log_2 \left(1 + \frac{h_{xy}}{N_0 + \sum_{t \in \phi^T \setminus \{y\}} h_{xt} l(||t - x||)} \right) \right], \qquad (2.3)$$

where h_{xy} and h_{tx} are independent random-variables representing the values of the fading power between the different transmitters and receiver and the expectation is with respect to this random vector of fades h. All of our theoretical results extend to this case but with a bit more notation and computation cost and thus, we discuss only the case without fading. The reason for fast-fading to not affect our theoretical insights is that both Equations (2.1) and (2.3) are deterministic monotone functions of the point x and ϕ. The rate functions are monotone in the sense that if $(x; y) \in \phi_1 \subseteq \phi_2$, then we have $R(x, \phi_1) \geq R(x, \phi_2)$ and $R^{(f)}(x, \phi_1) \geq R^{(f)}(x, \phi_2)$. We see from the proofs of our results, that these two (monotonicity and deterministic) are the crucial aspects of rate function on which the results hinge on and hence, we will only discuss the case without fading to simplify notation and convey the main ideas.

2.2.4 The Dynamics

This setup now allows one to precisely define the network dynamics. A link arriving with receiver in location $x_p \in \mathbf{S}$ and its transmitter at location

24

$y_p \in \mathbf{S}$ at time t_p with file of size L_p leaves the network at time d_p given by the following recursive definition

$$d_p = \inf \left\{ t > b_p : \int_{u=t_p}^{t} R(x_p, \phi_u) du \geq L_p \right\}. \tag{2.4}$$

In the above equation, ϕ_u denotes the point process of all links "alive" at time u i.e., $\phi_u^R = \sum_{p \in \mathbb{Z}} \delta_{x_p} \mathbf{1}_{\{u \in [b_p, d_p]\}}$ and $\phi_u^T = \sum_{p \in \mathbb{Z}} \delta_{y_p} \mathbf{1}_{\{u \in [b_p, d_p]\}}$ where δ_x denotes to the Dirac-measure at location $x \in \mathbf{S}$. We refer to the time instant b_p as the "birth" time of link p and d_p as the "death" time of link p. This is the justification for calling this dynamics a "spatial birth-death" model, i.e., this transmitter-receiver pair is "born" at time b_p and "dies" at time d_p and leaves the network.

This model is the wireless analog of the "flow-level" model introduced by Massoulie and Roberts [263] to evaluate and study wired networks, particularly the Internet. The flow-model in the present chapter is based on a more precise modeling of the wireless interactions compared to the standard conflict graph model of interference. This spatial birth-death model can also be viewed as a "dynamic" version of the model considered in [51], namely the Gaussian Interference channel with point-to-point codes. In our model, each link or a "flow" is a Guassian point-to-point channel using a point-to-point codebook and treats all Interference as Noise (IAN) as made evident in the rate-formulation in Equation 2.1. It was shown in [51], that one can consider other schemes such as Successive Interference Cancellation or Joint Optimal Decoding to get strictly better performance than considering Interference as

Noise in cases of static links that use ptp codes. We however only study the dynamic version of treating IAN and leave the other cases for future work.

2.2.5 Mathematical Assumptions

All the analysis and results rely on the following assumptions on the system model presented in the previous section.

1. The link arrival process is a time-space stationary Poisson Point Process of intensity λ. The probability of an arrival of a receiver in an infinitesimal location dx in an infinitesimal time interval dt is $\lambda dx dt$.

2. The file sizes of each transmitter are i.i.d. and exponentially distributed with mean L bits.

3. The transmitter location y of a receiver at x is assumed to be distributed uniformly and independently of everything else on the perimeter of a ball of radius T centered at x. In particular, the received signal power at any receiver is $l(T)$.

4. The thermal noise power $\mathcal{N}_0 > 0$ is a fixed constant.

5. The path-loss function is bounded and non-increasing with $l(0) = 1$. This is a reasonable assumption since energy is only dissipated on traveling through space and the received energy can be no larger than the transmit energy.

These assumptions (especially the statistical ones) are imposed primarily for mathematical tractability. It is well known, at least in the context of the Internet, that file sizes are Pareto [133] and it would make modeling sense to assume heavy-tailed file sizes. We will relax the statistical assumption on exponential file-sizes in the simulation studies. Nonetheless, studying the system under the Markovian statistical assumptions form a necessary first step before considering the general case.

In our model, we have that all links have the same length of T. This is commonly referred to as the 'Dipole-Model' of an ad-hoc wireless network [42]. An interesting limiting case is that of $T = 0$. This corresponds to the physical case of when the link lengths are very small compared to the size of the network. In this limiting case, the point process ϕ_t is simple and unmarked since the transmitter and receiver locations are identical, and the signal power is $l(0) = 1$. The interference function at a point x from configuration ϕ is then $I(x, \phi) = \sum_{y \in \phi \setminus \{x\}} l(||y - x||)$. We mention this limiting case here as it will help us to get a much better understanding of what our theoretical results imply, especially that of clustering (defined later in Definition 4). However, all of our mathematical results are valid for general arbitrary link distances T.

Although the assumptions may render the model somewhat specific, it still presents a formidable mathematical challenge and captures the key features of a spatio-temporal dynamic wireless network. Most prior works incorporating spatial interference circumvent this mathematical difficulty by making 'full-buffer' assumptions which is equivalent to assuming no temporal

interactions. Our results, especially the closed form expressions for approximating of delay are the first in the context of spatio-temporal wireless network models to the best of our knowledge.

The statistical assumptions, namely the Poisson arrival process and i.i.d. exponential file sizes imply that the process ϕ_t is a continuous time measure-valued Markov Chain on the state space of marked simple counting measure on $\mathbf{S}$ denoted as $\mathbf{M}(\mathbf{S})$ [136]. More precisely, the process ϕ_t is a piece-wise constant jump Markov Process i.e., from a time t, the next *change* in the configuration will occur after an exponentially distributed time duration with rate $\lambda|S| + \frac{1}{L}\sum_{x\in\phi_t} R(x,\phi_t)$. This interpretation follows since births occur at the epochs of an exponential clock with rate $\lambda|S|$ and the death rate of any receiver x in configuration ϕ is $\frac{1}{L}R(x,\phi)$ which is independent of everything else. The assumption $Q < \infty$ ensures that ϕ_t is a piece-wise constant jump process. Extending the analysis of stability to the case of $\mathbf{S} = \mathbb{R}^2$ is way more challenging and is left for future work. The large torus is meant to emulate the Euclidean space. The fact that it is similar to the Euclidean space (in terms of interference field and hence birth and death dynamics) justifies our use of the Palm calculus of the Euclidean space rather than that of the torus in some derivations.

The first natural question we ask about ϕ_t is that of time ergodicity which we address in the next section. Time ergodicity implies that the process ϕ_t admits an unique steady-state in which the links form a stationary and space-ergodic point process on $\mathbf{S}$. Moreover, since $\mathbf{S}$ is a compact set, the

stationary-regime when it exists will put only finitely many points in $\mathbf{S}$ at any given instant almost-surely. Denote by ϕ_0 the steady-state point-process of links i.e., the links that are "alive" or active in steady-state. ϕ_0 is a point-process on $\mathbf{S}$ with atoms representing the locations of receivers and marks representing the relative transmitter locations.

Denote by β the density of links present in the network in steady-state (assuming it exists). More formally, β denotes the intensity of the receiver point-process ϕ_0^{Rx} (which is the ground point process of ϕ_0) on $\mathbf{S}$ when the dynamics is in steady state. Note that the intensity of the transmitter point-process ϕ_0^{Tx} in steady-state is also β since every receiver in the model has exactly one transmitter. The distribution of the relative location of the transmitter of a typical receiver of ϕ_0^{Rx} is uniform on the perimeter of a ball of radius T around this receiver. However, the transmitter locations across different receivers of ϕ_0^{Rx} are not independent due to the correlation (clustering) induced by the dynamics.

The interpretation of time ergodicity is also connected to the phase-transition of mean delay. Little's law for this dynamics yields $\beta = \lambda W$, where W is the average sojourn time of a typical link i.e., $W = \mathbb{E}[d_0 - b_0]$; which follows from PASTA [387]. The process ϕ_t being time ergodic in our model is equivalent to asserting that $W < \infty$, i.e., finite mean delay for a typical link in the network. This interpretation is what we allude to in the system insight section which allows one to evaluate how frequently in space and time should the traffic arrival process be (i.e., how large λ) can be for the network

29

to provide finite mean-delay to all links.

2.3 Main Theoretical Results

The main theoretical results are on the time-ergodicity (or stability) conditions of the dynamics ϕ_t and on a certain structural characterization of the steady-state point process of ϕ_t whenever it exists. The proofs of the theorems are presented in the Appendix.

2.3.1 Stability Criterion

We state our main theoretical results on the stability criterion (i.e., time ergodicity) of the dynamics.

Theorem 1. If $\lambda > \frac{Cl(T)}{\ln(2)La}$, then the Markov Chain ϕ_t admits no stationary regime.

We see from the proof (in Section A.1) that this theorem only needs the weaker assumption that $l(\cdot)$ be such that $l(r) < \infty$ for all $r > 0$. This indeed is a weaker assumption than assuming that the function $l(\cdot)$ is bounded. Thus, we have as immediate corollary to this theorem:

Corollary 2. For the path-loss model $l(r) = r^{-\alpha}$, $\alpha \geq 2$, for all $\lambda > 0$, and all mean file sizes, the process ϕ_t admits no stationary-regime.

Proof. This follows since the integral $\int_{x \in \mathbf{S}} l(||x||)dx$ diverges for the function $l(r) = r^{-\alpha}$ for all $\alpha \geq 2$. $\qquad\square$

The next result provides a tight condition for time ergodicity.

Theorem 3. If $\lambda < \frac{Cl(T)}{\ln(2)La}$, then the Markov Chain ϕ_t is time ergodic, i.e., has an unique stationary regime.

Both of these Theorems are proved in Appendix A. We note that the above theorems statements are valid as is even in the case of fading if one used the rate-function in Equation (2.3) with the fades being unit-mean i.i.d. random variables. The two theorems identify the exact critical arrival rate λ for ergodicity as $\lambda_c = \frac{Cl(T)}{L\ln(2)a}$. We however refrain from studying the critical case as it is technically more subtle. In the sequel, whenever we refer to ϕ_0, we implicitly assume ϕ_t is ergodic, i.e., the condition $\lambda < \frac{C}{\ln(2)La}$ holds.

2.3.2 Clustering

In this section, we state the main structural characterization of the steady-state point process ϕ_0 when it exists i.e., when $\lambda < \frac{C}{L\ln(2)a}$. We need the following definition of clustering.

Definition 4. (CLUSTERING) Let ϕ be a stationary configuration of links, i.e., it is a stationary marked point-process on $\mathbf{S}$ with its marks in $\mathbf{S}$. Then ϕ is said to be clustered if for all bounded, positive, non-increasing functions $f(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$, the following inequality holds

$$\mathbb{E}^0_\phi[F(0,\phi)] \geq \mathbb{E}[F(0,\phi)], \tag{2.5}$$

where F is the shot-noise defined as follows. For any atom (receiver) $x \in \phi$

31

with its corresponding mark (transmitter) $y \in \mathbf{S}$, the shot noise $F(x, \phi) :=$ $\sum_{T \in \phi^{Tx} \setminus \{y\}} f(||T - x||)$.

Theorem 5. If the dynamics ϕ_t is ergodic, then the steady state point process ϕ_0 is clustered.

By substituting $f(\cdot) = l(\cdot)$ in Equation (2.5) , we get that the mean of the interference measured at any uniformly randomly chosen receiver in the steady-state point process (this is the interpretation of the Palm probability) is larger than the mean of the interference measured at any uniformly randomly chosen location of space in $\mathbf{S}$.

To understand why the above definition is a form of clustering, consider the case $T = 0$ which gives a clearer picture. In this case, Theorem (5) gives a clustering comparison of ϕ_0 with a Poisson Point Process (PPP) of same intensity. Let ψ be a PPP of the same intensity as ϕ_0. Then, from Slivnyak's theorem (Theorem 1.4.5, [42]), one can rewrite the inequality in (2.5) as

$$\mathbb{E}^0_{\phi_0}[F(0, \phi_0)] \geq \mathbb{E}^0_{\psi}[F(0, \phi_0)], \tag{2.6}$$

where $\mathbb{E}^0_{\psi}[F(0, \phi_0)] = \mathbb{E}[F(0, \psi)]$ follows from Slivnyak's theorem which is equal to $\beta \int_{x \in \mathbf{S}} f(||x||)dx$ from Campbell's Theorem (Theorem 1.4.3, [42]). Slivnyak's theorem essentially gives that the PPP has no clustering i.e., the Inequality 2.5 is an equality. Hence, we automatically have a shot noise comparison of the steady state point process ϕ_0 with a PPP.

The comparison with a PPP also gives us a comparison of the Ripley K-function [279] of ϕ_0 with that of a PPP. The Ripley K-function $K_\phi(\cdot)$:

$\mathbb{R}^+ \to \mathbb{R}^+$ of a point-process ϕ is defined as $K_\phi(r) = \frac{1}{\beta}\mathbb{E}_\phi^0[\phi(B(0,r)) - 1]$ where β is the intensity of ϕ and $\mathbb{E}_\phi^0$ is the Palm probability measure of ϕ. This function can be interpretted as the mean number of points (scaled by the intensity of the point-process) within distance r to the origin *conditioned* on a point of ϕ to be present at the origin. The Ripley K-function is commonly used in statistical analysis of point-patterns to identify if an empirical data-set exhibits statistical clustering [279]. Based on the shot-noise comparison with a PPP, we have the following corollary.

Corollary 6. Assume ϕ_t is in steady-state and $T = 0$. Denote by β to be the intensity of ϕ_0 and ψ to be a PPP on **S** with intensity β. Then, $K_{\phi_0}(r) \geq K_\psi(r)$.

Proof. Consider $f(x) = \mathbf{1}(x \leq r)$ in Theorem 5. □

We will use Ripley K-function in the simulations to compare the point process ϕ_0 with a PPP to derive a bound on the intensity β of ϕ_0 as a function of λ, L and $l(\cdot)$.

Intuitively, it is not surprising to expect a clustered point-process in steady state. An arriving link gets lower rate if it is in a crowded area of transmitters, due to interference. This arriving link also causes more interference to the cluster of links already present thereby causing more interference and slowing everyone down. This reinforcement of service slowdown is actually the fundamental reason making the system always unstable in the power law attenuation function case. More generally, when ϕ_t is sampled in steady-state,

it is expected to be clustered as formalized by Theorem 5. A snapshot of the point-process ϕ_0 is presented in Figure 2.4 which gives a visual illustration of the clustering.

2.4   Performance Analysis - Steady State Formulas

In this section, we propose two heuristic formulas for β the intensity of the point process ϕ_0 as a function of λ. Note that a heuristic formula for β gives a heuristic formula for mean delay W through Little's Law ($\beta = \lambda W$).

We propose two formulas - β_f called the *Poisson Heuristic* and β_s called *Second-Order heuristic* to approximate β the intensity of the steady-state point process ϕ_0. We show that subject to a natural conjecture (Conjecture 8), β_f is a lower bound on β. We see from simulations however that β_s is a much better approximation of β compared to β_f. Both formulas are derived based on approximately evaluating the following Equation which we establish in Equation (A.3) in the Appendix.

$$\lambda L = \beta \mathbb{E}^0_{\phi_0} \left[\log_2 \left(1 + \frac{l(T)}{\mathcal{N}_0 + I(0, \phi_0)} \right) \right]. \tag{2.7}$$

The Poisson Heuristic

The Poisson heuristic formula β_f is given by the largest solution to the following fixed point equation

$$\lambda L = \frac{\beta_f}{\ln(2)} \int_{z=0}^{\infty} \frac{e^{-\mathcal{N}_0 z}(1 - e^{-zl(T)})}{z} e^{-\beta_f q(z)} dz, \tag{2.8}$$

where $q(z) = \int_{x \in \mathbf{S}}(1 - e^{-zl(||x||)})dx$. This formula is obtained by approximating the expectation in Equation (2.7) by assuming the following "*Independent Poisson heuristic*". We assume that ϕ_0 is an independently marked Poisson-Point process with the transmitter locations of different receivers in ϕ_0 being independent. Since the transmitter locations are assumed to be independent, the process ϕ_0^{Tx} will also be a PPP in this Poisson heuristic. We state the following lemma without proof from [190] which is useful in computing the expectation under the Poisson assumption.

Lemma 7. Let X, Y be non-negative and independent Random Variables. Then,

$$\mathbb{E}\left[\ln\left(1 + \frac{X}{Y + a}\right)\right] = \int_{z=0}^{\infty} \frac{e^{-az}}{z}(1 - \mathbb{E}[e^{-zX}])\mathbb{E}[e^{-zY}]dz.$$

We can then explicitly compute the expectation in Equation (2.7) by letting $X = l(T)$ to be deterministic and $Y = I(0, \phi_0)$ as follows

$$
\begin{aligned}
\lambda L &= \beta_f \mathbb{E}_\psi^0 \left[\log_2\left(1 + \frac{l(T)}{\mathcal{N}_0 + I(0)}\right)\right] \\
&\overset{(a)}{=} \beta_f \mathbb{E}_\psi \left[\log_2\left(1 + \frac{l(T)}{\mathcal{N}_0 + I(0)}\right)\right] \\
&\overset{(b)}{=} \frac{\beta_f}{\ln(2)} \int_{z=0}^{\infty} \frac{e^{-\mathcal{N}_0 z}(1 - e^{-zl(T)})}{z}e^{-\beta_f q(z)}dz,
\end{aligned}
$$

where $q(z) = \int_{x \in \mathbf{S}}(1 - e^{-zl(||x||)})dx$ and ψ is a Poisson Point Process on $\mathbf{S}$ with intensity β_f. The equality (a) follows from Slivnyak's theorem and the

35

equality (b) follows from Lemma 7 and the formula for the Laplace functional of a Poisson Point Process. The subscript f refers to the computation of the density under this Poisson heuristic. This establishes the formula in Equation (2.8).

We now make the following conjecture on the higher-order moment measures of ϕ_0, which we will leverage to show that β_f is a lower bound on β.

Conjecture 8. Let ϕ_0 be the point process on $\mathbf{S}$ corresponding to the stationary distribution of ϕ_t with intensity β. Denote by ψ to be an independently marked Poisson Point Process on $\mathbf{S}$ with intensity β. The mark of any atom x of ψ is a point y drawn uniformly on the perimeter of a circle of radius T around x. Then, for any $s > 0$, we have $\mathbb{E}^0_{\phi_0}[e^{-sI(0;\phi_0)}] \leq \mathbb{E}^0_{\psi}[e^{-sI(0;\psi)}]$.

Note that from Slivnyak's theorem we also have $\mathbb{E}^0_{\psi}[e^{-sI(0;\psi)}] = \mathbb{E}_{\psi}[e^{-sI(0;\psi)}]$. This conjecture which is validated through simulations in Figure 2.1, is a slightly different statement on the structural characterization of ϕ_0 than stated in Theorem 5. This conjecture gives that the Laplace transform of the interference measured at a typical receiver of ϕ_0 is larger than that measured at a typical receiver of an equivalent PPP. In general, whenever we have ordering of the mean, then we have ordering of the Laplace Transform only as $s \to 0$. This ordering for the Laplace transform as $s \to 0$ follows from Taylor's expansion that $e^{-sx} \approx 1 - sx$ as $s \to 0$. However, in our case, we believe that the ordering on the Laplace transform holds for all $s \geq 0$ but we are unable to prove so. The intuition for this follows from the pictorial interpretation

36

that there are roughly the same number of interfering transmitters around a typical receiver in ϕ_0 and ψ since the intensities of ϕ_0 and ψ are the same. However, the interfering transmitters are closer to the typical receiver in ϕ_0 as compared to in ψ. This intuition follows from Corollary 6 where we had ordering of the Ripley-K function of ϕ_0 and ψ. This pictorial interpretation then gives an intuition for the conjecture since, the interference is the sum of attenuated powers from interfering transmitters where the attenuation is through a function that is non-increasing with distance. Thus, $I(0)$ is the sum of roughly the same number of terms in both ϕ_0 and in ψ, but each of the terms are slightly larger in ϕ_0 than in ψ. This interpretation can possibly be made rigorous in the asymptotic regime as $\lambda \uparrow \lambda_c$ by alluding to certain concentration phenomenon. However, we see from simulations that this conjecture holds true for all regimes of λ. This conjecture is further substantiated in Figure (2.2) which underpins Proposition 9.

The ordering of the mean does not always imply the ordering of Laplace transforms in general. As a very simple example consider two random variables X and Y where X takes values $\{1, 2, 3, 4\}$ with probabilities $\left\{\frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{3}\right\}$ and Y is deterministic and takes value of 2. Here $\mathbb{E}[X] = \frac{8}{3}$ and $\mathbb{E}[Y] = 2$. However, for $s = 1.1$, $\mathbb{E}[e^{-sX}] > \mathbb{E}[e^{-sY}]$. More generally if $\mathbb{E}[X] \geq \mathbb{E}[Y]$ but the higher order moments are ordered in the opposite direction, then one cannot expect an ordering on the Laplace-transform.

Proposition 9. Subject to Conjecture (8), we have that $\beta \geq \beta_f$, where β_f is the largest solution of Equation (2.8).
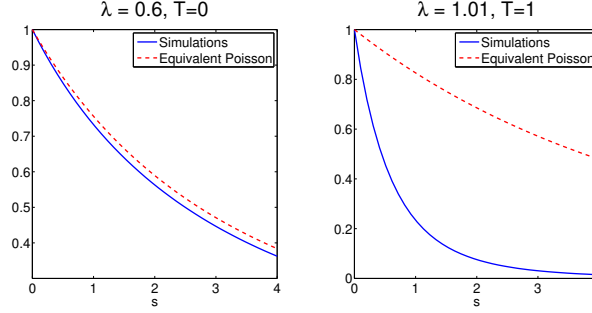
Figure 2.1: A plot comparing the functions $\mathbb{E}^0_{\phi_0}[e^{-sI(0;\phi_0)}]$ and $\mathbb{E}^0_{\psi}[e^{-sI(0;\psi)}]$, for $l(r) = (r+1)^{-4}$.

Proof. Let $g(\beta) = \beta\mathbb{E}^0_{\phi_0}[R(0;\phi_0)]$ (where ϕ_0 has intensity β) and let $p(\beta) = \beta\mathbb{E}^0_{\psi}[R(0;\psi)]$ where ψ is a PPP on $\mathbf{S}$ with intensity β. Rate-conservation equation (2.7) gives that $\lambda L = g(\beta)$ and our heuristic computation is $\lambda L = p(\beta_f)$. From our conjecture and Lemma 7, we have the inequality $g(\beta) \leq p(\beta)$. The function $g(\beta) = \beta\mathbb{E}^0_{\phi_0}[R(0;\phi_0)]$ is monotone non-decreasing in β as it describes the true dynamics through the equation $\lambda L = g(\beta)$. The monotonicity of $g(\cdot)$ along with the inequality $g(\beta) \leq p(\beta)$ gives the performance bound $\beta \geq \beta_f$. $\qquad\square$

Proposition 9 gives that $\beta_f|\mathbf{S}|$ is a lower bound on the mean number of links present in the network in steady state and $\frac{\beta_f}{\lambda}$, as a lower bound on mean-delay of a typical link.

The Poisson heuristic completely ignores the spatial clustering we established in Theorem 5 and assumes complete-spatial randomness. Since it does not account for the clustering it underestimates the typical interference seen at a receiver and therefore predicts a lower density of links. We see
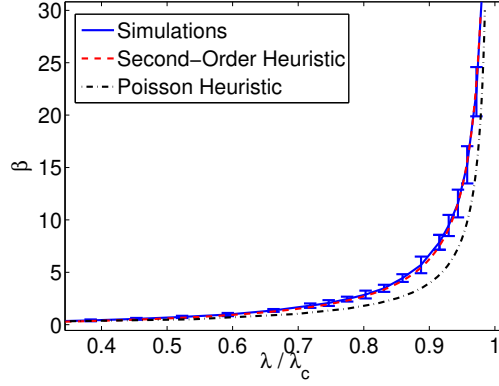
Figure 2.2: The performance plot with 95% confidence interval when $T = 0$ and $l(r) = (r + 1)^{-4}$.

through simulations, that this heuristic is poor (i.e., the gap between β and β_f is large) in certain traffic regimes (Figure 2.2). This is not surprising as one cannot neglect the effect of spatial correlations except in asymptotic regimes of heavy and light-traffic (detailed later). Motivated by the poor performance of the Poisson heuristic in certain regimes, we propose a "second-order heuristic" β_s which takes into account the spatial correlations by considering an approximation of the second-order moment measure of ϕ_0. We see through simulations (Figure 2.2) that β_s is a much better approximation of β than β_f in all traffic regimes.

Second-Order Heuristic

We propose a heuristic formula β_s for approximating β in Equation (2.9). For all values of T, β_s is given by

$$\beta_s = \frac{\lambda L}{C \log_2 \left(1 + \frac{l(T)}{\aleph_0 + I_s}\right)}, \tag{2.9}$$

where I_s is the smallest solution of the fixed-point equation

$$I_s = \lambda L \int_{x \in \mathbf{S}} \frac{l(||x||)}{C \log_2 \left(1 + \frac{l(T)}{\aleph_0 + I_s + l(||x||)}\right)} dx. \tag{2.10}$$

We call the heuristic in Equation (2.9) a *second-order heuristic* since it follows from an approximation of the second-order moment measure of ϕ_0 as follows. Let I_s denote the interference of a typical point at ϕ_0 and *assume* it is non-random and equal to its mean. Then, Equation (2.9) follows from Rate-Conservation in Equation (2.7). To compute I_s, we use the following approximation of the second order moment measure $\rho^{(2)}(x, y)$ of ϕ_0 as

$$\rho^{(2)}(x, y) \approx \frac{\beta \lambda L}{C \log_2 \left(1 + \frac{l(T)}{\aleph_0 + I_s + l(||x-y||)}\right)}. \tag{2.11}$$

Intuitively, the approximation is a form of cavity approximation which can be understood as follows. Two points at locations x and y will each "see" an interference of I_s which is the interference of a typical point plus the additional interference caused by the presence of the other point. Using the above interpretation of interference, Equation (2.11) is a form of Rate-Conservation on the pair of points at x and y. The average increase of the pair happens at

40

rate $2\lambda\beta$ and the average decrease of the pair happens at the rate equal to the sum of rates (since file-sizes are i.i.d. exponential) received by points x and y which is approximately $2(C/L)\log_2(1+\frac{l(T)}{N_0+I_s+l(||x-y||)})$ from the cavity approximation. Now, using the fact that $\mathbb{E}^0_{\phi_0}[I_0] := I_s = \frac{1}{\beta}\int_{x\in\mathbf{S}} l(||x||)\rho^{(2)}(x,0)dx$, we get Equation (2.10) from Equation (2.11).

The heuristic β_s is compared against the true β and the Poisson heuristic β_f in Figure 2.2. The second-order heuristic performs much better compared to the Poisson-heuristic as it takes into account some notion of spatial correlations which the Poisson heuristic completely ignores.

2.5   Simulation Studies

We perform simulations to gain a finer understanding of our model. We see that the bound in Proposition 9 is tight in the two asymptotic regimes of light and heavy-traffic where the effect of spatial correlations vanishes. We also argue that, in these two asymptotic regimes, the heuristic β_s is "close" to β_f thereby implying that β_s is also a good approximation to β. As noticed in Figure 2.2, β_f is a poor approximation to β compared to β_s in the intermediate traffic-regime which we further highlight in this section.

To explore the impact of spatial correlations, we study the tails of delay of a typical link and correlation between delays of different links in space. We observe that our model exhibits marked difference in terms of tail delay behavior from that of an equivalent queuing system which is obtained by a "spatial-fluid" approximation. We conclude from these studies that although

41

our model resembles that of a queue (for e.g. the dynamics satisfies Little's Law), there are significant differences due to the spatial correlations, which in hindsight is not so surprising. We finally perform simulations with heavy-tailed file size distribution and observe qualitatively the same phenomena as seen under exponential file-size distribution. We state our simulation results as claims which are not formal conjectures, but are meant to provide a starting point for future research.

2.5.1 Simulation Setup

The path-loss function we consider is $l(r) = (r + 1)^{-4}$. Although all of the results qualitatively hold for any bounded-non-increasing function, we choose this power law function due to its wide-spread popularity in modeling wireless propagation. We assume unit link-length $T = 1$ unless otherwise mentioned. We however note that all the qualitative results carry over for any value of T including the case of $T = 0$. The pictures of point-process and the Ripley K-functions we test are those corresponding to the receivers.

2.5.2 Tightness of $\beta_{\mathbf{f}}$

We study the bound in Proposition 9 by empirically noticing how much K_{ϕ_0}, the Ripley K-function of ϕ_0, deviates from that of an equivalent PPP denoted by K_{PPP}. The two Ripley K-functions being almost identical implies that the steady-state is "almost" Poisson and thereby the bound in Proposition 9 is good. On the other hand, if there is significant deviation between the two

Ripley K-function, then the bound is poor. We know from Corollary 6 that $K_{\phi_0}(r) \geq K_{\text{PPP}}(r)$ for all $r \geq 0$. Here, we are interested in seeing how large this difference can be.

To plot the Ripley K-functions, we simulated the Markov chain ϕ_t in forward time for a long time to obtain a single sample of the steady-state ϕ_0. We used the Spatstat package in R [60] to perform spatial statistics and plot K_{ϕ_0}. A single sample is sufficient as we take a large enough space (i.e., large $\mathbf{S}$) so that a single sample of ϕ_0 has about 500 points. Due to spatial ergodicity of ϕ_0, we get a smooth estimate of the K-function from a single sample.

We observe in Figures **??**, **??** and **??**, that the functions K_{ϕ_0} and K_{PPP} are very close in heavy and light traffic and are very different in intermediate traffic. The heavy traffic corresponds to the scenario when λ is very close to the critical λ_c and the light traffic corresponds to the case when λ is very "close" to 0. We do not rigorously demarcate the exact space-time scaling needed to define the two asymptotic limiting regimes as it is beyond the scope of this thesis.

Claim 10. ϕ_0 is almost Poisson in light-traffic. Moreover, the delay of a typical link converges weakly to an exponential distribution with mean $L \log_2 \left(1 + \frac{1}{N_0}\right)^{-1}$ as $\lambda \to 0$.

In the light-traffic regime, λ is very "small" compared to L, and thereby β is also "small". This then implies that the distribution for the interference $I(0, \phi_0)$ is close to 0, thereby making the interaction between the points al-

most negligible. The rate-function can then be approximated as $R(x, \phi) \approx \log_2\left(1 + \frac{1}{N_0}\right)$ and the dynamics resembles that of a spatial $M/M/\infty$ queue whose stationary spatial distribution is a PPP. The intensity β in this regime is $\beta_l = \lambda \left(\log_2\left(1 + \frac{1}{N_0}\right)\right)^{-1}$. The subscript l refers to the density computation in the interaction-less approximation. Figure **??** provides numerical evidence that ϕ_0 exhibits very little clustering in this regime and is "close" to a PPP.

Claim 11. In the heavy-traffic regime, ϕ_0 is almost Poisson, i.e., the effect of clustering vanishes as $\lambda \to \lambda_c$.

The intuition behind the heavy-traffic behavior is that as λ approaches λ_c, the stationary distribution is very dense, i.e., β is large. Hence, the interference of a typical arriving link is mainly dominated by the local geometry which does not change much during the life-time of the typical link. This indicates that the dynamics behaves very similarly to a heavily loaded $M/M/1$ Processor Sharing (PS) queue and the correlation across space is negligible in this regime. Moreover, it is easy to see that $\lambda = \lambda_c$ is an asymptote for Equation (2.8) i.e., as $\lambda \to \lambda_c$, $\beta_f \to \infty$. This further strengthens the belief that the stationary distribution is close to Poisson in the heavy-traffic regime as it predicts the correct stability boundary. Making this claim rigorous or even just state a mathematical conjecture is quite challenging and would require an appropriate scaling of space and time similar to the diffusion scaling considered for a single server PS queue [181].
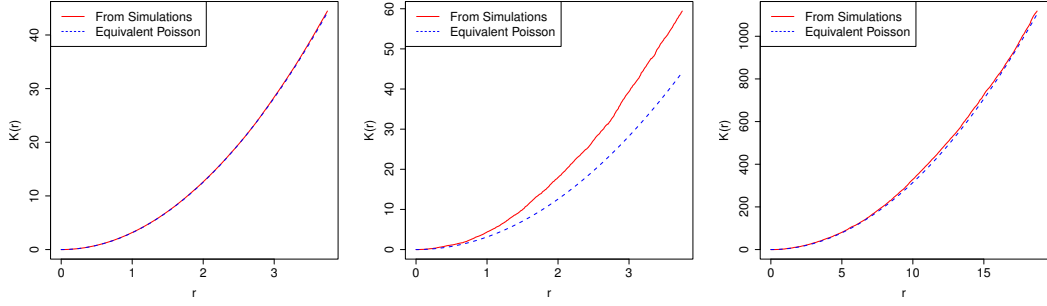
Figure 2.3: Plot comparing the Empirical Ripley K-function K_{ϕ_0} with that of an equivalent PPP. The path loss function is $l(r) = (r+1)^{-4}$, $T = 1$. The critical $\lambda_c = 1.42$. This shows that there is little clustering in the heavy and light traffic regimes but significant clustering in the intermediate regime.

2.5.3 Tightness of β_s

We argue here that in both the low and heavy-traffic regimes, the approximation β_s is close to β_f and is hence a good approximation of the true β. In low-traffic, as $\lambda \to 0$, the smallest solution of Equation (2.10) tends to 0 and hence the formula for $\beta_s \approx \dfrac{\lambda L}{C \log_2\left(1 + \frac{1}{N_0}\right)}$. This from Claim 10 gives that β_s and β_f predict the same value in low-traffic. In high-traffic regime, as $\lambda \to \lambda_c$, the value of I_s from Equation (2.10) is very high. Thus, the second-order moment-measure approximation $\rho^{(2)}(x, y)$ in Equation (2.11) is almost constant i.e., does not depend of the actual values of x and y as $l(\cdot)$ is a bounded function. This implies that the effect of clustering vanishes in this heuristic and hence is close to β_f.

45

2.5.4 Intermediate Clustered Regime

In the intermediate regime, the Poisson approximation is poor and the steady state-point process is quite clustered (see Figures 2.4 and **??**) i.e., K_{ϕ_0} is much larger than K_{PPP}. However, we see from Figure 2.2 that the second-order heuristic β_s performs much better than the Poisson heuristic in this regime as it takes into account some form of spatial-correlations atleast upto second-order moment measure of ϕ_0. However, Figure 2.4 which shows a snapshot of ϕ_0 which is a clustered process is very interesting as it indicates finer properties of higher order moment measures. One observes for instance "*filaments*" of points which are locally directional in-spite of the fact that the dynamics is isotropic. Such behavior indicates that the higher order moment measures of ϕ_0 (of order greater than 2) may have interesting properties which we capture neither in Theorem 5 nor in the second order heuristic β_s. Understanding the higher order moment measure of ϕ_0 can also aid in proposing a *provably* better performance bound in this intermediate regime. Studying these higher order moment measures of ϕ_0 will be a very interesting and challenging direction of research.

2.5.5 Delay Tails

To get a heuristic understanding of the delay tails, one would be tempted at first glance to approximate our model by an equivalent $M/M/1$ PS queue using a *spatial-fluid* approximation that neglects randomness in space. We see through simulations that any approximation that neglects spatial interactions
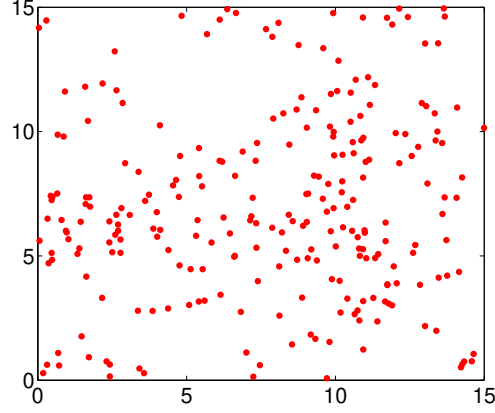
Figure 2.4: A sample of ϕ_0 when $\lambda = 0.99$ and $l(r) = (r+1)^{-4}$. This is a visual representation of the clustering of points.

will predict much larger delays for a typical link than the true delays in our model.

Claim 12. The delay tails in our model are exponential and have a faster decay than that of an equivalent $M/M/1$ PS queue obtained by a "spatial-fluid" approximation.

An equivalent $M/M/1$ PS queue approximation has the following parameters - arrival rate λ, service requirement of mean L and service capacity of the server λ_c which is split equally among all customers in the queue. Such a PS queuing model is equivalent to a first-order approximation where the spatial randomness vanishes and a point in steady-state receives rate of $C \log_2 \left(1 + \frac{1}{N_0 + \beta a} \right)$ where β is the density of points in steady-state. Hence, the quantity $\frac{C}{\ln(2)a}$ (which is an upper bound on the total rate given to all points i.e., $C\beta \log_2 \left(1 + \frac{1}{N_0 + \beta a} \right) \leq \frac{C}{\ln(2)a}$) can be seen as the maximum service

47

capacity of the spectrum in $\mathbf{S}$ which is equally shared by all links accessing the spectrum. Another simple picture as to why the above $M/M/1$ PS queue is a simple heuristic is to observe that this queue corresponds to the scenario when one ignores spatial interactions among the arriving points and assumes that the total spectrum "capacity" of λ_c is shared equally among all the links sharing the spectrum in $\mathbf{S}$. Hence, the mean-delay under the $M/M/1$ - PS model for a typical point is $\frac{L}{\lambda_c - \lambda}$ and the stability criteria for this queue is the same as that for our spatial model. However, we note from simulations (Figure 2.6) that the delay tails predicted by the heuristic $M/M/1$ queue which completely ignores spatial interactions are much larger than those observed in our model.

The poor performance of the queuing approximation can be understood by studying the correlation between the delays of different links. In Figure 2.7, we plot the correlation between the delay experienced by two links arriving at the *same time* as a function of their distance. We consider the $T = 0$ case and hence the distance between two links is just the distance between the two points. Numerically, we plotted Figure 2.7 by first sampling a steady-state point process (by running the Markov Chain ϕ_t for a long time) and then introducing two additional links to this sample with independent file-sizes. We then run the dynamics from this state until the two additional links die and then compute the correlation between their delays

We see from Figure 2.7 that as the distance between the two links

increases, the delays of the two links are almost uncorrelated even though they arrive at the same time. This indicates that, two links arriving at the same time will be almost oblivious to each other and will each roughly receive independent service if they arrive far enough apart in space. This is unlike in the $M/M/1$ - PS queue approximation where two customers arriving at the same time have positively associated delays as both of them will be competing for the same spectrum resource. This suggests that the spatial heterogeneity is key in extracting more "service" from the spectrum than predicted by a model which considers spectrum as a fixed quantity of good to be divided among contending links.
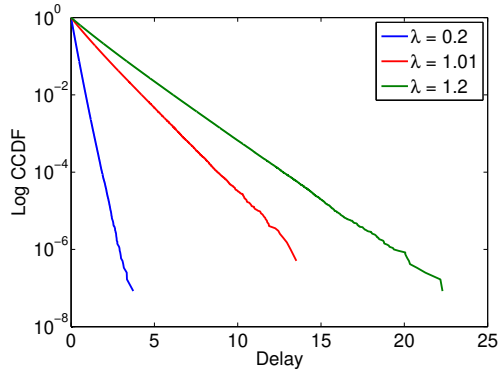
Figure 2.5: Plot of logarithm of CCDF of delay.

2.5.6 Heavy Tailed File Sizes

Claim 13. ϕ_t, with file-sizes being Pareto distributed of mean L and finite variance, admits a stationary regime with the critical λ being smaller than or equal to $\frac{C}{\ln(2)La}$.

49
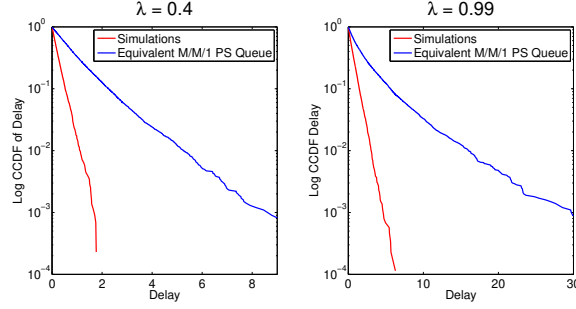
Figure 2.6: Comparison of the delays with that of an equivalent M/M/1 - PS queue. The critical $\lambda_c = 1.42$.
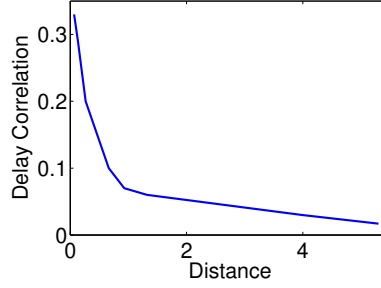
Figure 2.7: Decay of delay correlation of two points born at the same time, as a function of their distance. $\lambda = 0.8$.

This model also exhibits the interesting phenomenon of prominent clustering in the intermediate traffic regime and very little to no-clustering in the asymptotic regimes of high and low traffic. Note that the term "high-traffic" in this context is somewhat loose since we do not even know exactly the stability region. With regards to delays, our model predicts tails that are stochastically dominated by the delay of a typical customer of an equivalent $M/GI/1$ PS queue (see Figure 2.8). The equivalent queue we compared against had a capacity of λ_c which from Claim 13 is an upper bound on the capacity. Nonetheless, the delay predicted in our model is stochastically smaller. This

50

observation again highlights the importance of taking into account the spatial heterogeneity in modeling the "service" provided by the spectrum.
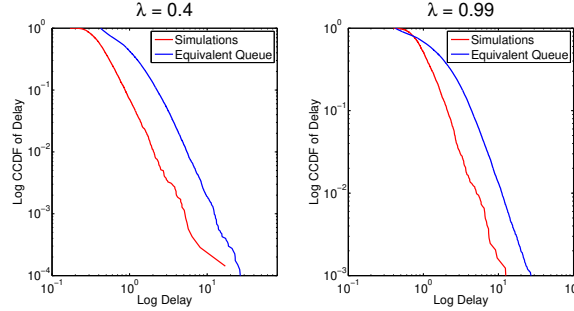
Figure 2.8: Comparision of the delay under Pareto file size distribution with mean L and shape $\alpha = 2.5$.

2.6   Conclusion and Future Work

In this chapter, we proposed a novel space-time interacting particle system to model spectrum sharing in ad-hoc wireless networks. We computed exactly the phase-transition point for time ergodicity. We also proved the intuitive fact that the steady-state point-process corresponding to this dynamics exhibits clustering. In order to understand the performance metric of density of links in steady-state, we proposed a Poisson heuristic β_f (which is a bound subject to Conjecture 8) and a second order heuristic β_s. We saw from simulations that both the heuristics are tight in the two asymptotic regimes of heavy and light traffic. However, in the intermediate traffic regime, we found that the heuristic β_s performs much better compared to the Poisson heuristic β_f as β_s accounts for some spatial correlations which are non negligible in this regime. We also saw through simulations that any form of simplistic model-

ing of spatio-temporal interactions through PPP or equivalent queues ignoring spatial clustering, leads to poor estimates for performance.

From a mathematical perspective, we identified several challenging directions of future work in the simulation section. In particular, understanding the higher order moment measure of ϕ_0 will be key in evaluating or providing provably tighter bounds for performance metrics. Understanding the higher-order moment measures may also aid in making progress on Conjecture 8. From an information-theoretic perspective, we considered a dynamic interference network where links treat interference as noise. However, it will be interesting to consider other receiver schemes such as Successive Interference Cancellation or Joint-Decoding and show that the critical arrival rate for these schemes are strictly better than considering all Interference as Noise. This will then yield the complete dynamic version of the model considered in [51], namely a dynamic version of an interference network with point-to-point (ptp) codes. In this thesis, we carry out the program started here in the case when the spatial domain is infinite. This is particularly interesting both mathematically, as it will require the development of new tools and ideas and also from an engineering perspective

Chapter 3

Interference Queueing Networks on Grids

3.1 Introduction

In this chapter, we continue the line of inquiry started in the previous chapter on the spatial-birth death model for a wireless network. We refer the reader to Chapter 2 for practical and applied motivations based on emerging wireless networks to study such spatial birth-death processes. In the previous chapter, we showed that when the arrival rate is sufficiently small, the network, which forms a Markov process on the space of all marked point processes on a finite torus $\mathbf{S}$ admits a stationary regime. The key factor that was used throughout in our analysis was that the spatial domain $\mathbf{S}$ was a compact subset of the Euclidean plane. Thus, we could stochastically dominate the Markov process above and below by finite dimensional Markov chains, and use the classical theory of such Markov processes to reason about the behaviour of our spatial birth death dynamics. However, the results and analysis presented in that chapter does not answer if the dynamics instead were defined on an infinite domain, (i.e., $\mathbf{S} = \mathbb{R}^d$), then what would be its behaviour, and in particular, does it also admit a stationary regime for λ smaller than the identified critical

---

Parts of this chapter is published in [338]. The author was part of formulating, executing and writing up the results in that paper.

arrival rate. The main motivation of study in this chapter thus is to advance our understanding of the spatial birth-death wireless network dynamics on infinite domains.

To this effect, we introduce a discretized version of the spatial birth-death model, where the spatial domain is the integer grid. However, we crucially consider the case when the spatial domain is infinite, i.e., the spatial domain is the infinite grid. Recall from the proof of Theorem 3, the behaviour of the dynamics on a discrete grid is in a precise sense a good approximation of the dynamics on the continuum. Hence, we decided to focus on the discrete space setting in this chapter. In future work, we plan to extend the ideas developed in this chapter to study the continuum space case. From a practical viewpoint, one may wonder what modeling benefits does a network model comprising of an infinite space bring. Our motivations to study an infinite network model are two fold. First, from a mathematical point of view, the infinite model provides an elegant method of modeling dynamics on classical stochastic geometric objects, which are typically defined on an infinite domain. The infinite domain also presents us with certain algebraic symmetries such as translation invariance which greatly aid in understanding the dynamics both qualitatively and quantitatively. We see in all the proofs that translation invariance plays a very key role. Second and more importantly in our context, the infinite network provides a clean mathematical abstraction to understand many phenomena pertaining to large finite networks. Just as in the Ising Model [241] for example, in our case, the infinite model could poten-

tially admit multiple stationary distributions (See Theorem 25 and the remark following it). If it is the case, by analogy to the Ising model, it could indicate the presence of long range spatial correlations. Practically, operating the network in this 'phase' could be undesirable. Thus understanding the infinite model could pave the way for more refined analysis of large finite networks, as done for other interacting particle systems such as the Ising model. In view of this, we find it important to describe and analyze the infinite network model along with large finite networks.

3.1.1 Organization of the Chapter

We describe the model in Section 3.2 and state the main results of this chapter in Section 3.3. For brevity, we skip presenting the proofs of all the results and refer the reader to [338] for the proofs. Instead, we highlight the main technical challenges and provide ideas and proof sketches in Section 3.4 and Chapter B. We survey related work in Section 3.5 and conclude the chapter in Section 3.6.

3.2 Mathematical Framework and the Network Model

Formally, we consider a spatial queueing network consisting of an infinite collection of processor sharing queues interacting with each other in a translation invariant way. In our model, there is a queue located at each grid point of $\mathbb{Z}^d$, for some $d \geq 1$. The queues evolve in continuous time and serve the customers according to a generalized processor-sharing discipline.

The arrivals to the queues form a collection of i.i.d. Poisson Point Processes of rate $\lambda > 0$. Thus, the total arrival rate to the network is infinite since there is an infinite number of queues. The different queues interact through their departure rates. We model the interactions through an *interference sequence* that we denote by $\{a_i\}_{i \in \mathbb{Z}^d}$. It is such that $a_i \geq 0$ and $a_i = a_{-i}$ for all $i \in \mathbb{Z}^d$. We also assume that this sequence is finitely supported, i.e., $L := \max\{||i||_\infty : a_i > 0\} < \infty$. For ease of exposition, we also assume that $a_0 = 1$ in certain sections of the chapter, although our model and its analysis can be carried out for any non-zero value of a_0. For any $t \in \mathbb{R}$, let $\{x_i(t)\}_{i \in \mathbb{Z}^d} \in \mathbb{N}^{\mathbb{Z}^d}$ denote the queue lengths at time t in the network, i.e., the state of the system at time t. Then *the interference experienced by a customer located in queue i at time t* is defined as $\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)$, i.e., some weighted sum of queue lengths of the *neighbors* of queue i. Observe that the neighborhood definition is translation invariant. Conditional on the queue lengths $\{x_i(t)\}_{i \in \mathbb{Z}^d}$ at time t, the instantaneous departure rate from any queue i at time t is given by $\frac{x_i(t)}{\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)}$, with $0/0$ interpreted as being equal to 0. Note that since the interference sequence $\{a_i\}_{i \in \mathbb{Z}^d}$ is non-negative, and $a_0 = 1$, for all $t \in \mathbb{R}$ and all $i \in \mathbb{Z}^d$, the instantaneous departure rate from queue i at time t is $\frac{x_i(t)}{\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)} \in [0, 1]$ and is hence bounded. Since $\{a_i\}_{i \in \mathbb{Z}^d}$ is non-negative, the rate of service at a queue is reduced if its 'neighbors' have larger queue lengths. This is meant to capture the fundamental spatiotemporal dynamics in wireless networks where the instantaneous rate of a link is reduced if there are a lot of other links accessing the spectrum nearby, due to an increase of in-

56

terference. In the rest of the chapter, we shall always assume that there exists at least one $i \in \mathbb{Z}^d \setminus \{0\}$ such that $a_i > 0$. For otherwise, the system is 'trivial', as the queues evolve independent of each other without any interaction amongst them, according to a standard $M/M/1$ dynamics with unit service rate. Observe that the Markovian dynamics of our model is non-reversible and does not fall under the class of generalized Jackson networks. This model is also not of the mean-field interacting queues type such as the supermarket model [380], which admit a form of 'asymptotic independence' across queues, as the system sizes get large.

The model described above can be viewed as a space discretized version of the model described in Chapter 2. The grid points of $\mathbb{Z}^d$ in the present model represent tiny 'chunks' of continuum space of $\mathbb{R}^d$. The number of customers at some time t in queue $z \in \mathbb{Z}^d$ is the number of links at time t located in the tiny region of space corresponding to queue z. For simplicity, we assume that links are very tiny (mathematically a point) and hence both the transmitter and receiver of a link can be seen to be located at the same point. This is of-course only done to have a clear presentation and convey the main message and one can easily have non-zero link lengths as done in [336], with significantly heavier notation. This is the key difference between our model and those of for example [345], where a queue represents a link in the model of [345]. In contrast, in our model, a queue represents a region of space, and the queue length denotes the number of transmitting devices in that region of space. We assume that, the interference function $l(\cdot)$ is discretized as the

sequence $\{a_i\}_{i \in \mathbb{Z}^d}$. Thus a_i represents the received power at queue 0 (or queue i) if unit power is transmitted at queue i (or at queue 0). However, we relax the condition that a_i be non-increasing in $||i||$ but impose the stricter condition that it is finitely supported. Thus the total interference at any link in queue 0 at time t is $a_0(x_0(t) - 1) + \sum_{i \in \mathbb{Z}^d \setminus \{0\}} a_i x_i(t)$. To derive the instantaneous rate, we assume a linear approximation of the Shannon rate and model the instantaneous rate of communication is given by **SINR**. Thus, the instantaneous rate of communication at any link in queue 0 at time t is then given by $\left(a_0(x_0(t) - 1) + \sum_{i \in \mathbb{Z}^d \setminus \{0\}} a_i x_i(t) + N_0 \right)^{-1}$, where N_0 is the thermal noise power. The minus 1 for the a_0 term is to account for the fact that a link will not interfere with itself. For further ease of notation, we assume $N_0 = a_0$ and thus obtain that the instantaneous rate of communication for any link at queue 0 is given by $\frac{1}{\sum_{i \in \mathbb{Z}^d} a_i x_i(t)}$. Now as there are $x_0(t)$ links at queue 0 and they all have independent exponential unit mean file sizes, the total rate of departure is then given by $\frac{x_0(t)}{\sum_{j \in \mathbb{Z}^d} a_j x_j(t)}$.

To summarize, the model in this chapter differs from that presented in Chapter 2 in three key aspects. First, the space in our model is discretized, while it is a continuum in Chapter 2. This is not much a difference as already seen in the previous chapter in the proof of Theorem 3. Second, the rate of departure of a link in our model is given just by the SINR, as opposed to the Shannon's formula used in the previous chapter. This can be viewed as a *low SINR* regime, or high noise regime. From a mathematical perspective however, this simplification turns out to be very frutiful and allows us to derive closed

form formulas as explained in the sequel. Thirdly, the model considered in this chapter refers to the infinite spatial domain while the model in the previous chapter corresponded to only compact spatial domains.

We now give a precise description of our model in subsection 3.2.1 and demonstrate certain useful monotonicity properties it satisfies in Subsection 3.2.3. We then precisely state the definition of stability in Section 3.2.4 and the notion of stationary solutions to the dynamics in Section 3.2.5.

3.2.1 Probabilstic Framework

Our model is parametrized by $\lambda \in \mathbb{R}$ and an *interference sequence* $\{a_i\}_{i \in \mathbb{Z}^d}$ which is a non-negative sequence. This sequence satisfies $a_0 = 1$, $a_i = a_{-i}$ for all $i \in \mathbb{Z}^d$ and $L := \sup\{||i||_\infty : a_i > 0\} < \infty$, i.e., is finitely supported. We also impose the sequence $\{a_i\}_{i \in \mathbb{Z}^d}$ to be *irreducible*, which gives that for all $z \in \mathbb{Z}^d$, there exists $k \in \mathbb{N}$ and $i_1, \cdots, i_k \in \mathbb{Z}^d$ not necessarily distinct, such that $i_1 + i_2 \cdots + i_k = z$ and $a_{i_j} > 0$ for all $j \in \{1, \cdots, k\}$. To describe the probabilistic setup, we assume there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that contains the stationary and ergodic driving sequences $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d}$. For each $i \in \mathbb{Z}^d$, $\mathcal{A}_i$ is a Poisson Point Process (PPP) of intensity λ on $\mathbb{R}$, independent of everything else and $\mathcal{D}_i$ is a PPP of intensity 1 on $\mathbb{R} \times [0, 1]$, independent of everything else. Our stochastic process denoting the queue lengths $t \to \{x_i(t)\}_{i \in \mathbb{Z}^d}$ will be constructed as a factor of the process $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d}$. The process $\mathcal{A}_i := \sum_{q \in \mathbb{Z}} \delta_{A_q^{(i)}}$ encodes the fact that, at times $\{A_q^{(i)}\}_{q \in \mathbb{Z}}$, there is an arrival of a customer in queue i. Thus the arrivals to queues form

59

PPPs of intensity λ and are independent of everything else. The process $\mathcal{D}_i := \sum_{q \in \mathbb{Z}} \delta_{(D_q^{(i)}, U_q^{(i)})}$ encodes that there is a possible departure from queue i at time $D_q^{(i)}$, with an additional independent $U[0,1]$ random variable provided by $U_q^{(i)}$. To precisely describe the departures, we define *the interference at a customer in queue i at time t* as equal to $\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)$. A customer, if any, is removed from queue i at times $D_q^{(i)}$ if and only if $U_q^{(i)} \leq \frac{x_i(D_q^{(i)})}{\sum_{j \in \mathbb{Z}^d} a_{j-i} x_j(D_q^{(i)})}$. In other words, conditionally on the state of the network $\{x_j(D_q^{(i)})\}_{j \in \mathbb{Z}^d}$ at time $D_q^{(i)}$, we remove a customer from queue i at time $D_q^{(i)}$ with probability $\frac{x_i(D_q^{(i)})}{\sum_{j \in \mathbb{Z}^d} a_{j-i} x_j(D_q^{(i)})}$, independently of everything else. Thus we see that conditionally on the network state $\{x_j(t)\}_{j \in \mathbb{Z}^d}$ at time t, the instantaneous rate of departure from any queue $i \in \mathbb{Z}^d$ at time $t \in \mathbb{R}$ is $\frac{x_i(t)}{\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)}$, independently of everything else. Observe that since $a_0 = 1$, if $x_i(t) > 0$, then necessarily, $\frac{x_i(t)}{\sum_{j \in \mathbb{Z}^d} a_j x_{i-j}(t)} \in (0,1]$.

We further assume (without loss of generality) that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with a group $(\theta_u)_{u \in \mathbb{R}}$ of measure preserving functions from Ω to itself where θ_u denotes the 'time shift operator' by $u \in \mathbb{R}$. More precisely $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d} \circ \theta_u$ is the same driving sequence where each of the arrivals and departures are shifted by time u in all queues, i.e., if $\mathcal{A}_i := \sum_{q \in \mathbb{Z}} \delta_{A_q^{(i)}}$ and $\mathcal{D}_i := \sum_{q \in \mathbb{Z}} \delta_{(D_q^{(i)}, U_q^{(i)})}$, then $\mathcal{A}_i \circ \theta_u := \sum_{q \in \mathbb{Z}} \delta_{A_q^{(i)} - u}$ and $\mathcal{D}_i \circ \theta_u := \sum_{q \in \mathbb{Z}} \delta_{(D_q^{(i)} - u, U_q^{(i)})}$, for all $i \in \mathbb{Z}^d$. We also assume that the system $(\mathbb{P}, (\theta_u)_{u \in \mathbb{R}})$ is ergodic, i.e., if for some event $A \in \mathcal{F}$, if $\mathbb{P}[A \triangle A \circ \theta_u] = 0$ for all $u \in \mathbb{R}$, then $\mathbb{P}[A] \in \{0, 1\}$.

3.2.2 Construction of the Process

Before we analyze the above model, one needs to ensure that it is 'well-defined'. We mean that our model is well defined if given the initial network state $\{x_i(0)\}_{i \in \mathbb{Z}^d}$, any time $T \geq 0$ and any index $k \in \mathbb{Z}^d$, we are able to construct the queue length $x_k(T)$ unambiguously and exactly. In the case of finite networks (i.e., networks with finitely many queues), the construction is trivial: almost surely, one can order all possible events in the network with increasing time, and then update the network state sequentially using the evolution dynamics described above. Such a scheme works unambiguously since, almost surely, all event times will be distinct and in any interval $[0, T]$, there will be finitely many events. The main difficulty in the case of infinite networks is that there is no *first-event* in the network. In other words, in any arbitrarily small interval of time, infinitely many events will occur almost surely and hence we cannot construct by ordering all the events in the network. However we show in Appendix B, that in order to determine the value of any arbitrary queue $k \in \mathbb{Z}^d$ at any time $T \geq 0$, we can effectively restrict our attention to an almost surely finite subset $X_{k,T} \subset \mathbb{Z}^d$ and determine $x_k(T)$ by restricting the dynamics to $X_{k,T}$ to the interval $[0, T]$. This is then easy to construct as it is a finite system. Thereby we can determine $x_k(T)$ unambiguously. The construction techniques we use are common in Interacting Particle systems setup (for example, the book of [241] for more examples and details on the method).

3.2.3 Monotonicity

We establish an obvious but an extremely useful property of path-wise monotonicity satisfied by the dynamics. Note that our model is not monotone separable in the sense of [53] since the dynamics does not satisfy the external monotonicity condition. Nonetheless, the model still enjoys certain restricted forms of monotonicity, which we state below. We only highlight the key idea for the proof and defer the details to Appendix B.

Lemma 14. If we have two initial conditions $\{x_i'(0)\}_{i\in\mathbb{Z}^d}$ and $\{x_i(0)\}_{i\in\mathbb{Z}^d}$ such that for all $i \in \mathbb{Z}^d$, $x_i'(0) \geq x_i(0)$, then there exists a coupling such that $x_i'(T) \geq x_i(T)$ for all $i \in \mathbb{Z}^d$ and all $T \geq 0$ almost surely.

The proof is by a path-wise coupling argument, where the two different initial conditions are driven by the same arrival and potential departures. The key idea is as follows. At arrival times, the ordering will trivially be maintained. Consider some queue i and time t where there is a potential departure. If $x_i'(t) \geq x_i(t) + 1$, then, since at most one departure occurs, the ordering will be maintained. But if $x_i'(t) = x_i(t)$, then the rates $\frac{x_i'(t)}{\sum_{j\in\mathbb{Z}^d} a_j x_{i-j}'(t)} \leq \frac{x_i(t)}{\sum_{j\in\mathbb{Z}^d} a_j x_{i-j}(t)}$ and hence the ordering will again be maintained. This observation can be leveraged again to have the following form of monotonicity.

Lemma 15. For all initial conditions $\{x_i(0)\}_{i\in\mathbb{Z}^d}$, for all $0 \leq s \leq t \leq \infty$, all $X \subset \mathbb{Z}^d$, and all $T > 0$, $\{x_i(T)\}_{i\in\mathbb{Z}^d}$ is coordinate-wise larger in the true

dynamics than in the dynamics constructed by setting $\mathcal{A}_j([s,t]) = 0$ for all $j \in X$.

3.2.4 Stochastic Stability

We establish a $0 - 1$ law stating that either all queues are transient or all queues are recurrent (made precise in Lemma 16 in the sequel). Thus, we can then claim that the entire network is stable if and only if any (say queue indexed 0 without loss of generality) is stable (made precise in Definition 17 in the sequel). To state the lemmas, we set some notation. Let $T \geq 0$ and $s > -T$ be arbitrary and finite. Denote by $\{x_{i;T}(s)\}_{i \in \mathbb{Z}^d}$ the value of the process seen at time s when started with the empty initial state at time $-T$, i.e., with the initial condition of $x_{i;T}(-T) = 0$ for all $i \in \mathbb{Z}^d$. Lemma 14 implies that for every queue $i \in \mathbb{Z}^d$, and for $\mathbb{P}$ almost-every $\omega \in \Omega$, we have $T \to x_{i;T}(s)$ is non-decreasing for every fixed s. Thus, for every i, and every $s \in \mathbb{R}$, there exists an almost sure limit $\lim_{T \to \infty} x_{i:T}(s) := x_{i;\infty}(s)$. From the definition, this limit is shift-invariant, i.e., almost surely, for all $x \in \mathbb{R}$, we have $x_{i;\infty}(s) \circ \theta_x = x_{i;\infty}(s + x)$.

Lemma 16. We have either $\mathbb{P}[\cap_{i \in \mathbb{Z}^d}\{x_{i;\infty}(0) = \infty\}] = 1$ or $\mathbb{P}[\cap_{i \in \mathbb{Z}^d}\{x_{i;\infty}(0) < \infty\}] = 1$.

The proof follows from standard shift-invariance arguments which we present here for completeness. Since for all $x \in \mathbb{R}$ and all $j \in \mathbb{Z}^d$, $x_{i;\infty}(0) \circ \theta_x = x_{i;\infty}(x)$, we have that this lemma implies for all $s \in \mathbb{R}$, either $\mathbb{P}[\cap_{i \in \mathbb{Z}^d}\{x_{i;\infty}(s) = \infty\}] = 1$ or $\mathbb{P}[\cap_{i \in \mathbb{Z}^d}\{x_{i;\infty}(s) < \infty\}] = 1$.

Proof. It suffices to first show that for any fixed $i \in \mathbb{Z}^d$, we have $\mathbb{P}[x_{i;\infty}(0) < \infty] \in \{0,1\}$. Assume that we have established for some i (say 0 without loss of generality that) $\mathbb{P}[x_{0;\infty}(0) < \infty] \in \{0,1\}$. From the translation invariance of the dynamics, it follows that, for all $i \in \mathbb{Z}^d$, we have $\mathbb{P}[x_{i;\infty}(0) < \infty] = \mathbb{P}[x_{0;\infty}(0) < \infty]$. Thus, if $\mathbb{P}[x_{0;\infty}(0) < \infty] = 1$, then $\mathbb{P}[\cap_{i \in \mathbb{Z}^d} x_{i;\infty}(0) < \infty] = 1$. Similarly, if $\mathbb{P}[x_{0;\infty}(0) = \infty] = 1$, then $\mathbb{P}[\cap_{i \in \mathbb{Z}^d} x_{i;\infty}(0) = \infty] = 1$. Thus to prove the lemma, it suffices to prove that $\mathbb{P}[x_{0;\infty}(0) < \infty] \in \{0,1\}$.

The key observation is, the event $A := \{\omega \in \Omega : x_{0;\infty}(0) < \infty\}$ is such that for all $x \in \mathbb{R}$, $\mathbb{P}[A \triangle A \circ \theta_u] = 0$. To show this, first notice that from elementary properties of PPP, we have that for every $i \in \mathbb{Z}^d$ and every compact set $B \subset \mathbb{R}$, $\mathcal{A}_i(B) < \infty$ a.s.. Now for any $x \geq 0$, we have $x_{0;\infty}(0) \circ \theta_x \leq x_{0;\infty}(0) + A_0([0,x])$, which is finite almost surely if $x_{0;\infty}(0) < \infty$ almost surely. Similarly for every $x < 0$, $x_{0;\infty}(0) = x_{0;\infty}(0) \circ \theta_x + A_0([x,0])$, which again implies that $x_{0;\infty}(0) \circ \theta_x$ is almost surely finite if $x_{0;\infty}(0) < \infty$. Thus, for all $x \in \mathbb{R}$, we have $\mathbb{P}[A \triangle A \circ \theta_x] = 0$, which from ergodicity of $(\mathbb{P}, (\theta_u)_{u \in \mathbb{R}})$ implies $\mathbb{P}[A] \in \{0,1\}$ and thus the lemma is proved. $\qquad\square$

The following definition of stability follows naturally.

Definition 17. The system is **stable** if $x_{0;\infty}(0) < \infty$ almost surely. Conversely, we say the system is **unstable** if $x_{0;\infty}(0) = \infty$ almost surely.

Observe that the definition of stability does not require $\mathbb{E}[x_{0;\infty}(0)]$ to be finite. In words, we say that our model is stable if when starting with all queues being empty at time $-t$ in the past, the queue length of any queue stays

64

bounded at time 0 when letting t go to infinity. This definition of stability is similar to the definition introduced for example by [251] in the single server queue case. A nice account of such backward coupling methods can be found in [49].

The main result in this chapter is to prove that if $\lambda \sum_{j \in \mathbb{Z}^d} a_j < 1$, then the system is stable (Theorem 19). Moreover, in this case, we compute exactly the mean queue length in steady state, i.e., an explicit formula for $\mathbb{E}[x_{0;\infty}(0)]$, and by shift-invariance it is equal to $\mathbb{E}[x_{i;\infty}(s)]$. We also conjecture this condition to be necessary, i.e., if $\lambda \sum_{j \in \mathbb{Z}^d} a_j > 1$, then $x_{0;\infty} = \infty$ almost surely. We are unable to prove this conjecture yet, but prove it for the special case of $d = 1$ in Theorem 28.

3.2.5 Translation Invariant Stationary Solutions

Definition 18. A probability measure $\boldsymbol{\pi}$ on $(\mathbb{Z}^d)^{\mathbb{N}}$ is said to be translation invariant, if $(y_i)_{i \in \mathbb{Z}^d} \sim \boldsymbol{\pi}$ implies, for all $x \in \mathbb{Z}^d$, $(y_{i-x})_{i \in \mathbb{Z}^d} \sim \boldsymbol{\pi}$. A probability measure $\boldsymbol{\pi}$ on $(\mathbb{Z}^d)^{\mathbb{N}}$ is said to be stationary for the dynamics $\{x_i(t)\}_{i \in \mathbb{Z}^d}$ if, whenever $\{x_i(0)\}_{i \in \mathbb{Z}^d}$ is distributed according to $\boldsymbol{\pi}$ independently of everything else, then, for all $t \geq 0$, the random variables $\{x_i(t)\}_{i \in \mathbb{Z}^d}$ are also distributed as $\boldsymbol{\pi}$.

In this chapter, we restrict ourselves to studying stationary solutions to the dynamics that are translation invariant in space. Observe that the driving sequence $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d}$ is translation invariant on $\mathbb{Z}^d$, i.e., for all $v \in \mathbb{Z}^d$, $(\mathcal{A}_{i-v}, \mathcal{D}_{i-v})_{i \in \mathbb{Z}^d}$ is equal in distribution to $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d}$. Furthermore, the in-

teractions among the queues are also translation invariant, since the definition of interference seen at a queue is translation invariant. However, it is not immediately clear that all stationary solutions must necessarily be translation invariant. It is known, for instance in the literature on Ising Models (see the book [176]), that certain stationary measures for translation invariant Glauber dynamics need not necessarily be translation invariant. We leave the question of existence and construction of non-translation invariant stationary measures for our model to future work.

Moreover, as our network is not finite-dimensional, stability in the sense of Definition 17 does not imply ergodicity in the usual Markov chain sense. In particular, it does not imply that stationary distributions are unique, and starting from any initial condition on $\mathbb{N}^{\mathbb{Z}^d}$, the queue lengths converge in some sense to the minimal stationary distribution considered in Definition 17. Stability only implies the *existence* of a stationary solution, namely the law of $\{x_{i\infty}(0)\}_{i\in\mathbb{Z}^d}$. However, uniqueness is not granted and one of our main results in Proposition 20 bears on this. Moreover, convergence to stationary solutions from different starting states is more delicate as evidenced in Theorems 23 and 25.

3.3   Main Results

Our first result provides a sufficient condition for stability. Moreover, it precisely gives a closed form expression for the mean queue length.

Theorem 19. If $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then the system $\{x_i(\cdot)\}_{i \in \mathbb{Z}^d}$ is stable. Furthermore, for all $i \in \mathbb{Z}^d$ and $s \in \mathbb{R}$, the minimal stationary solution $\{x_{i;\infty}(s)\}_{i \in \mathbb{Z}^d}$ satisfies

$$\mathbb{E}[x_{i;\infty}(s)] = \frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j}.$$

In the rest of the chapter, the condition $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$ will be referred to as the stability criterion for the system. In this theorem, we only considered whether there exists a stationary solution to the dynamics. However, as our network consists of infinitely many queues, uniqueness of stationary solutions is not guaranteed. As explained, we are mainly concerned with stationary solutions of queue lengths that are translation invariant in space. The following Proposition sheds light on the question of unique translation invariant stationary solutions.

Proposition 20. If $\mathbb{E}[x_{0;\infty}(0)^2] < \infty$, then $\{x_{i;\infty}(0)\}_{i \in \mathbb{Z}^d}$ is the unique translation invariant stationary solution with finite second moment.

This result relies on the finiteness of second moment of the stationary queue length, which does not follow immediately from the conclusions of Theorem 19. In this regard, we have the following proposition, that establishes finiteness of second moment under further restrictive conditions than stability.

Proposition 21. If $\lambda < \frac{2}{3} \frac{1+c}{\sum_{j \in \mathbb{Z}^d} a_j}$, where $c = \frac{\sqrt{a_0^2 + a_0 \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j} - a_0}{\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j}$, then we have $\mathbb{E}[x_{0;\infty}(0)^2] < \infty$.

Note that under our assumption of $a_0 = 1$, the value of the constant c can be simplified as $c = \frac{\sqrt{\sum_{j \in \mathbb{Z}^d} a_i - 1}}{\sum_{j \in \mathbb{Z}^d} a_i - 1}$. Observe that if $c = \frac{1}{2}$, then the above proposition will cover the full range of stability. However, for any valid interference sequence $\{a_i\}_{i \in \mathbb{Z}^d}$, we have $c \in \left(0, \frac{1}{2}\right)$, with $c \nearrow \frac{1}{2}$ as $\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j \searrow 0$. Thus, this proposition does not cover the full stability region. For the simplest non-trivial case of one dimensions and the interference sequence being $a_i = 1$ if $|i| \leq 1$ and $a_i = 0$ if $|i| > 1$, the second moment is finite for $\lambda \leq 0.91 \frac{1}{\sum_{j \in \mathbb{Z}} a_j}$. From Propositions 20 and 21, we have the following immediate corollary.

Corollary 22. If $\lambda < \frac{2}{3} \frac{1+c}{\sum_{j \in \mathbb{Z}^d} a_j}$, where c is given in Proposition 21, then $\{x_{i;\infty}(0)\}_{i \in \mathbb{Z}^d}$ is the unique translation invariant stationary solution with finite second moment.

Our next set of results assesses whether queue length process converges to any stationary solution when started from different starting states. Observe that we deemed the system stable if when started with all queues empty, the queue lengths converge to a proper random variable. Thus, stability alone does not imply convergence from other initial conditions. In this regard, our main results are stated in Theorems 23 and 25 which show the sensitivity of the dynamics to the starting conditions. In particular, we show in Theorem 23, that if λ is sufficiently small and the initial conditions are uniformly bounded, then the queue lengths converge to the minimal stationary solution. Surprisingly, in Theorem 25 below, we exhibit both deterministic and random initial conditions *for all* $\lambda > 0$, such that the queue lengths diverge, even though the

68

stability criterion $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$ is met. This is a new type of result which holds primarily since the network consists of an infinite collection of queues.

Theorem 23. Let λ be such that the minimal stationary solution satisfies $\mathbb{E}[x_{0;\infty}(0)^2] < \infty$. Then if the initial condition satisfies $\sup_{i \in \mathbb{Z}^d} x_i(0) < \infty$, the queue length process $\{x_i(\cdot)\}_{i \in \mathbb{Z}^d}$ converges weakly to the minimal stationary solution as $t \to \infty$.

As the queue lengths are positive integer valued, and the dynamics admits a form of monotonicity, every fixed finite collection of coordinates also converges to the minimal stationary solution in the total variation norm in the above Theorem, which is stronger than just weak convergence. Notice from Proposition 21, that if $\lambda < \frac{2}{3} \frac{1+c}{\sum_{j \in \mathbb{Z}^d} a_j}$, where c is given in Proposition 21, then the conclusion of the above Proposition holds.

We further examine sensitivity to initial conditions in Theorem 25 by constructing examples where the queue lengths diverge, even though the stability criterion is met. To state the result, we need a natural 'irreducibility' condition on the interference sequence $\{a_i\}_{i \in \mathbb{Z}^d}$.

Definition 24. The interference sequence $\{a_i\}_{i \in \mathbb{Z}^d}$ is said to be *irreducible* if, for all $z \in \mathbb{Z}^d$, there exists $k \in \mathbb{N}$ and $i_1, \cdots, i_k \in \mathbb{Z}^d$, not necessarily distinct, such that $i_1 + i_2 \cdots + i_k = z$ and $a_{i_j} > 0$ for all $j \in [1, k]$.

This is a natural condition which ensures that we cannot 'decompose' the grid into many sets of queues, each of which does not interact with the

queues in the other group. In the extreme case, this disallows the case when $a_i = 0$ for all $i \neq 0$, in which case the network can be decomposed into an infinite collection of independent $M/M/1$ Processor Sharing queues.

Theorem 25. For all $\lambda > 0$, $d \in \mathbb{N}$, and irreducible interference sequences $\{a_i\}_{i \in \mathbb{Z}^d}$, and even when the stability criterion holds, there exists

1. A deterministic sequence $(\alpha_i)_{i \in \mathbb{N}}$ such that if the initial condition satisfied $x_i(0) \geq \alpha_i$ for all $i \in \mathbb{Z}^d$, then the queue length of 0 satisfies $\lim_{t \to \infty} x_0(t) = \infty$ almost surely.

2. A distribution ξ on $\mathbb{N}$ such that if the initial condition $\{x_i(0)\}_{i \in \mathbb{Z}^d}$ is an i.i.d. sequence with each $x_i(0)$, $i \in \mathbb{Z}^d$ being distributed as ξ independent of everything else, then the queue length of 0 (or any finite collection of queues) satisfies $\lim_{t \to \infty} x_i(t) = \infty$ almost surely.

Based on the proof of this theorem, we make the following remark.

Remark 26. For all λ, the support of $(\alpha_i)_{i \in \mathbb{Z}^d}$ in statement 1 above can be made arbitrarily sparse, i.e. for any sequence $(b_n)_{n \in \mathbb{N}}$ such that $b_n \to \infty$, the initial conditions $(\alpha_i)_{i \in \mathbb{Z}^d}$ can be chosen, such that $\lim_{n \to \infty} \frac{\sum_{i \in \mathbb{Z}^d : ||i||_\infty \leq n} \mathbf{1}_{\alpha_i > 0}}{b_n} = 0$, yet the queue lengths converge almost surely to infinity.

The above theorem is qualitative in nature, as it only establishes the existence of bad initial conditions, but does not provide estimates for how large this initial condition must be. In this regard, we include Proposition 27,

which pertains to the deterministic starting state in the simplest non-trivial system, namely the case of $d = 1$, and the interference sequence being $(a_i)_{i \in \mathbb{Z}}$ such that $a_i = 1$ for $|i| \leq 1$ and $a_i = 0$ otherwise. This simplest non-trivial example already contains the key ideas and hence we present the computations involved explicitly here. In principle, one can provide a quantitative version of the above theorem in full generality. However, we do not pursue this here as they involve heavy calculations without additional insight into the system.

Proposition 27. Consider the system with $d = 1$ and the interference sequence $a_i = 1$ if $|i| \leq 1$ and $a_i = 0$ otherwise. Let $(b_n)_{n \in \mathbb{N}}$ be arbitrary deterministic non-negative integer valued sequence such that $b_n \to \infty$. If the initial condition, $\alpha_i := i^{i2^{i+2}+8}$ for $i \in \{b_n : n \in \mathbb{N}\}$, and $\alpha_i = 0$ otherwise, then for every $\lambda > 0$, $\lim_{t \to \infty} x_0(t) = \infty$ almost surely.

Regarding the converse to stability, we prove the following result in Theorem 28, which establishes that the phase-transition at the critical λ is sharp, at least in certain cases, and we conjecture it to be sharp for all cases.

Theorem 28. For the dynamics on $d = 1$ and monotone interference sequence $\{a_i\}_{i \in \mathbb{Z}}$, if $\lambda > \frac{1}{\sum_{j \in \mathbb{Z}} a_i}$, then there exists a N_0 sufficiently large so that for all $N \geq N_0$, the dynamics truncated to the finite set $[-N, \cdots, N]$ is transient.

Observe that the truncated dynamics to a finite set is a classical finite dimensional Markov Chain. From the monotonicity in the dynamics, we have the following immediate corollary.

71

Corollary 29. For the system with $d = 1$ and monotone interference sequence, if $\lambda > \frac{1}{\sum_{j \in \mathbb{Z}} a_i}$, then the system is unstable.

3.3.1 Open Questions and Conjectures

We now list some conjectures and questions that are left open by the present chapter. The first one concerns the moments of the minimal stationary solution $x_{0;\infty}(0)$. We have established an exact formula for the mean that holds in the entire stability region and finiteness of the second moment in a fraction of the stability region. In an earlier version of this work that we posted online, we had put forth the following conjecture.

Conjecture 30. If $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then $\mathbb{E}[x_{0;\infty}(0)^2] < \infty$.

Subsequently, this has been proven to be correct by [356] using rate-conservation techniques, similar to those presented in the present chapter. This fact, along with Proposition 20 implies, that the minimal stationary solution is indeed the unique translation invariant stationary solution to the dynamics that admits finite second moments. Furthermore, the conclusion of Theorem 23 also hold for all $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, namely from all bounded initial conditions, the queue length process will converge to this unique translation invariant stationary solution if $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$. In this regard, three natural interesting questions arise - one concerning what other moments of stationary queue lengths are finite, one regarding correlation decay and another on existence of other stationary solutions.

Question 31. For each $\lambda \in \left(0, \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}\right)$, what moments of $x_{0;\infty}(0)$ are finite ?

In an ongoing project [339], we show that for all $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, there exists a $c > 0$ such that $\mathbb{E}[e^{cx_{0;\infty}(0)}] < \infty$. In particular, this will establish that all power moments of $x_{0;\infty}(0)$ are finite

Question 32. How does the correlation $k \to \mathbb{E}[x_{0;\infty}(0)x_{k;\infty}(0)] - \mathbb{E}[x_{0;\infty}(0^2)]$ decay as $|k| \to \infty$?

Question 33. Does the dynamics admit stationary solutions other than the minimal one ? If so, do there exist initial conditions such that the law of the queue lengths converge to them ?

We know from Proposition 20 that the minimal stationary solution is the unique translation invariant stationary solution with finite second moment. This then raises the following question.

Question 34. Does there exist a translation invariant stationary solution that has an infinite first moment ? Does there exist one with finite first moment, but infinite second moment ?

In regard to establishing transience, a natural open question in light of Theorem 28 is to extend this result to higher dimensions and non monotone interference sequence. We make the following conjecture.

Conjecture 35. For all $d \geq 1$ and interference sequence $(a_i)_{i \in \mathbb{Z}}$, if $\lambda > \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then the system is unstable.

3.4 Main Ideas in the Analysis

The key technical challenge in analyzing our model is the positive correlation between queue lengths, which persist even in the model with infinitely many queues (see also Figure 3.1). As mentioned, our system of queues is neither reversible, nor falls under the category of generalized Jackson networks. Thus, our model does not admit a product form stationary distribution, even when there are finitely many queues. In particular, the model has no asymptotic independence properties as those encountered in "mean-field models" (such as the supermarket model [380]). The correlations across queues is intuitive, since if a queue has a large number of customers, then its neighboring queues will receive lower rates, and thus they will in turn build up. Therefore in steady state, if a particular queue is large, most likely, its neighboring queues are also large (see also Figure 3.1).

To prove the sufficient condition for stability, we first study finite space-truncated torus systems. In words, we restrict the dynamics to a large finite set $B_n \subset \mathbb{Z}^d$, and study its stability by employing fluid-like and Lyapunov arguments. This model in a sense is the discrete space analog of the compact continuum birth-death process studied in Chapter 2. For this model, we write down rate conservation equations in Section B.2 and solve for the mean queue-length of this dynamics. The equations of rate-conservation contains the key technical innovations in this model. The rate conservation equations turn out to be surprisingly fruitful, as we are able to obtain an exact formula for the

mean queue length. This formula also gives as a corollary, that the queue length distributions are tight, as the size of the truncation B_n increases to $\mathbb{Z}^d$. We then show that we can take a limit as B_n increases to all of $\mathbb{Z}^d$ and consider the stationary solution $\{x_{i;\infty}(0)\}_{i \in \mathbb{Z}^d}$ as an appropriate limit of the stationary solutions of the space-truncated system. The central argument in this section is to exploit the many symmetries, the monotonicity of the dynamics and the aforementioned tightness to arrive at the desired conclusion. We furthermore, apply a similar rate conservation equation for the infinite system, which along with monotonicity arguments, establishes the uniqueness of stationary solutions with finite second moments.

To study the convergence from different initial conditions, we employ different arguments, again exploiting the symmetry and monotonicity in the model. To show that stability implies convergence from bounded initial conditions, we define a modified K-shifted system. It is a model having the same dynamics as our original model, except that the queue lengths do not go below K, for some $K \in \mathbb{N}$. We carry out the same program of identifying a bound on the first moment on the minimal stationary solution to the shifted dynamics by analyzing similar rate conservation equations as for the original system. We then exploit the monotonicity and the fact that a stationary solution with finite mean is unique, to conclude that stability implies convergence to the minimal stationary solution from bounded initial conditions. In order to identify initial conditions from where the queue length can diverge even though the stability condition holds, we first consider a simple idea of 'freez-

ing' a boundary of queues at a large distance n, to a 'large value' α_n around a typical queue, say 0, and then consider its effect on the queue length at the origin. By freezing, we mean, there are no arrivals and departures in those queues, but a constant number α_n of customers that cause interference. We see that by choosing α_n sufficiently large, this *wall* can influence the stationary distribution at queue 0. We leverage this observation, along with monotonicity, to construct both deterministic and random translation invariant initial conditions such that queue lengths diverge to $+\infty$ even though the stability condition holds. This proof technique is inspired by similar ideas developed to establish non-uniqueness of Gibbs measures in the case when the state space of a particle is finite, while our methods and results bear on the case when the state space is countable.

3.5   Related Work

Our study is motivated by the performance analysis of wireless networks which has a large and rich literature as already outlined in Chapter 2 (see for ex. [86] [42], [360] and the references therein). Thus, we survey the related work in this discrete setting and place it in context mainly with the literature in applied mathematics where similar models have appeared in other contexts. In particular, since some of the new properties we establish in this chapter are directly linked to the fact that there are infinitely many queues, we thought it appropriate to briefly survey the mathematical literature on queueing models consisting of infinitely many queues interacting through some translation in-

variant dynamics. A model related to ours is the so called *Poisson Hail* model which has been studied in a series of papers [52],[50],[164]. The discrete version of this model consists of a collection of queues on $\mathbb{Z}^d$, where the queues interact through their service mechanism in a translation invariant manner. In this model, the customer at a queue occupies a 'footprint' and when being served, no other customer in the queues belonging to its footprint is served. In contrast, in our model a customer slows down the customers in neighboring queues, but does not block them. Another set of papers close to our model is [48], [259], and [257]. These papers analyze an infinite collection of queues in series. The main results are connections with last passage percolation on grids. A similar model to this is studied by [160], where analogues of Burke's theorem are established for a network of infinite collection of queues on the integers. There is also a series of papers on infinite polling systems. The paper [165] considers a polling model with an infinite collection of stations, and addresses questions about ergodicity and positive recurrence of such models. In a similar spirit, [93] considers infinite polling models and establishes the presence of many stationary solutions leveraging the fact that the Markov process is not finite-dimensional. The dynamics in these polling systems are however very different from ours. The paper of [258] also introduced a nice problem with translation invariant dynamics, but only analyzed the setting with finitely many queues. The paper of [211] introduced an elegant problem on Jackson queueing networks on infinite graphs. However, the stationary distribution there admits a product-form representation, which is very different

from our model in the present paper. The paper [189] studies translation-invariant dynamics on infinite graphs arising from combinatorial optimization, which again falls broadly in the same theme, but for a fundamentally different class of problems. Queueing like dynamics on an infinite number of nodes are also studied, though under different names, in the interacting particle system literature in the sense of [241]. The most well known instance of interacting particle system connected to queueing is probably the TASEP. Another fundamental class of interacting particle system exhibiting a positive correlation between nodes (like our model) is the ferromagnetic Ising model. The first difference is that the state-space of a node is not compact (i.e., $\mathbb{N}$, since the state is the number of customers in the queue) in our model, whereas it is finite in these models. Another fundamental difference between our model and these is the lack of reversibility. The common aspects are the infinite dimensional Markovian representation of the dynamics, the non uniqueness of stationary solutions, and the sensitivity to initial conditions. Infinite queueing models are also central in mean-field limits. In the literature on mean-field queueing systems ([380],[180],[147]) the finite case exhibits correlations among the queue lengths thereby making them difficult to analyze. However, in the large number of node limit, one typically shows that there is 'propagation of chaos'. This then gives that the queue lengths become independent in the limit. This independence can then be leveraged to write evolution equations for the limiting dynamics which can be analyzed. Such mean-field analysis have recently become very popular in the applied literature (for ex. [392],[25]). Our model
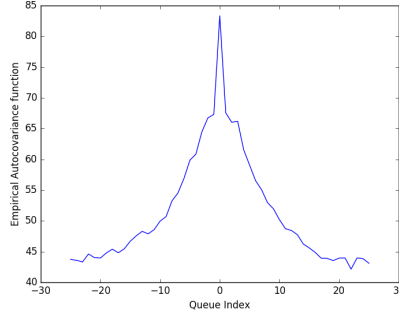
Figure 3.1: A plot of the empirical covariance function of queue lengths in steady state. We consider $d = 1$ and 51 queue placed on a ring. The arrival $\lambda = 0.1419$ while $\lambda_c = 1/7$ and the interaction function is $a_i = 1$ if $|i| \leq 3$ and 0 otherwise.

differs fundamentally from the above models in many aspects. First, unlike the mean-field models described above, we can directly define the limiting infinite object, i.e., a model with infinitely many queues. Secondly and more crucially, our infinite model does *not* exhibit any independence properties in the limit, i.e., queue lengths are positively correlated even in the infinite model. This is why we need different techniques to study this model. Our main technical achievement in this context is to introduce coupling and rate conservation techniques not relying on any independence properties.

3.6 Conclusions

In this chapter, we introduce a novel model of infinite spatial queueing system with the queues interacting with each other in a translation invariant fashion. This model was motivated by scalability questions of operating wireless networks, that was raised in the previous chapter. In particular, we answer

in the affirmative, that in case of finitely supported interference sequence, the qualitative phase transition result in the previous chapter also holds for the infinite network case. Nevertheless, the mathematical arguments needed to establish this result is very different from those that were needed in the previous chapter. The queueing model we introduced here is neither reversible nor admits any mean-field type approximations to analyze it. In the present chapter, we analyzed this model using rate conservation and coupling arguments, which can be of interest to study other large interacting queueing systems. We establish a sufficient condition for stability which we also conjecture to be necessary. Surprisingly, we are able to compute an exact formula for the mean number of customers in steady state in any queue. Furthermore, we identify a subset of the stability region in which the stationary solution with finite second moment is unique. Interestingly however, we see that our system is sensitive to initial conditions. We construct for every λ, both a deterministic and translation invariant random initial conditions, such that the queue lengths diverges to infinity almost surely, even though the stability conditions hold.

However, our work leaves open many intriguing questions as discussed in Section 3.3.1. In particular, the correlation across queues is interesting as it can be numerically simulated and is shown in Figure 3.1. In Figure 3.1, we are empirically estimating the function $i \to \mathbb{E}[(x_0(t)-\mu)(x_i(t)-\mu)]$, where μ is the mean queue length given in the formula in Theorem 19. However, we cannot

simulate an infinite system and hence consider a finite system of 51 queues placed on a ring (i.e., one dimensional torus). We use the interaction function $a_i = 1$ if $|i| \leq 3$ and 0 otherwise. The critical arrival rate is 0.14285 and we used a $\lambda = 0.1419$ to simulate. The mean queue length in this example is 21.18. We estimate the function $\mathbb{E}[(x_0(t) - \mu)(x_i(t) - \mu)]$ by collecting many independent samples approximating the steady state queue lengths $\{x_i^{(25)}\}_{i \in [-25,25]}$. For each collected sample, we evaluate an empirical covariance function by setting the value at $i \in [-25, 25]$ to be $(x_i^{(25)} - \mu)(x_0^{(25)} - \mu)$, where μ is the mean queue length equal to 21.18 in this example. We plot after averaging over many such functions computed on independent queue-length samples. From the plot, the strong positive correlations are very evident, as the function plotted is always large and positive. The figure also supports our intuition that the correlations must decay with distance as one would guess, but yields no concrete insight for the exact nature of this decay, for instance does the correlations decay polynomially or exponentially with the distance. Exploring this and other related questions in our model is an exciting line of future work.

Chapter 4

User Association in Cellular Networks

4.1 Introduction

This chapter studies the problem of user association in modern day large cellular networks. In traditionally operated mobile networks, each user (mobile) is obliged to subscribe to a particular operator and has access to the base stations owned by the operator (or to Wi-Fi access points administered by the operator). A new paradigm known as *mobile virtual network operators* (MVNO) is currently reshaping the wireless service industry. The idea is to provide higher service quality and connectivity by pooling and sharing the infrastructure of multiple wireless networks. A concrete example of this is Google's new "Project Fi" service in the United States, whose main feature is improved coverage provided through outsourcing infrastructure from its partners, T-Mobile, Sprint and their Wi-Fi networks. In parallel, the European Commission has been ruling favorably for MVNOs since 2006, so as to make the European wireless market more competitive [14], thereby facilitating investment in MVNOs in Europe. These virtual operators can take advantage of the different physical network operators which use separate bandwidths, and

---

Parts of this chapter is published in [340]. The author was part of formulating, executing and writing up the results in that paper.

even different wireless access technologies, for improvement of user experience. It is reported [15] that the market share of these virtual operators, especially in mature markets, ranged from 10% (UK and USA) to 40% (Germany and Netherlands) as of 2014. This increase of diversity, both in terms of bandwidths and wireless technologies raises many challenging questions to realize the envisioned benefits in large scale networks. In this context, this chapter considers the design and analysis of user association schemes, that are both scalable to large systems and leverage the gains from diversity.
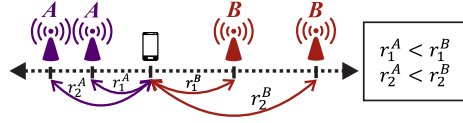
Figure 4.1: Motivational example with two technologies: $r_1^A < r_1^B$ and $r_2^A < r_2^B$.

The most widely studied association policy in existing wireless networks consists in associating each user equipment (UE) with the nearest base station (BS) or access point where one typical aim is to maximize the likelihood of being covered or connected. One of the main points of the present chapter is that this is no longer optimal in these emerging virtual networks, as further discussed in Section 4.1.2. The subtle distinction arising from technology diversity is illustrated in Fig. 4.1, where r_1^A and r_2^A (respectively, r_1^B and r_2^B) denote the distances to the nearest and second-nearest BSs of technology A (respectively, technology B) from the UE located at the origin. Also, we assume that $r_1^A < r_1^B$, i.e. the nearest BS of technology A is the nearest to the

UE, and there are only four BSs as shown in Fig. 4.1, which are identical except that they operate on different technologies (i.e. non-overlapping bandwidths). In the single technology case ($A = B$), the UE can simply associate with the BS at r_1^A. However, if $A \neq B$, e.g. the two technologies operate on different bandwidths, the locations of the strongest interferers, r_2^A and r_2^B (the second-nearest BSs), may *overturn* the choice of technology A when the strongest interferer of technology B is much farther from the UE than that of technology A, i.e. $r_2^B \gg r_2^A$, thus boosting the signal-to-noise-ratio (SINR) of technology B. In light of this example, optimal association in such networks requires new paradigms. We can further generalize the above example and envisage a practical scenario where each UE can obtain the information about several received pilot signals of nearby BSs, as in 3G and 4G cellular networks, which can be translated into a vector of distances. In this chapter, we are interested in investigating the following question.

Q: *How to leverage the diversity brought in by the choice to associate with various non-overlapping networks ? How much should a mobile phone learn about the instantaneous network conditions to exploit the diversity ?*

4.1.1 Main Contributions

Our main contribution in this chapter is to formalize the above question into an abstract mathematical framework. We propose a simple abstraction for the network model and quantify the question how much resources should

a mobile phone expend in learning about the instantaneous network condition. Secondly, we look at the question of given that the UE has expended its resources to have an accurate knowledge of the instantaneous network conditions, how should it exploit this information to improve performance.

4.1.2 Related Work

The policy of associating each UE to the nearest BS or the BS with the strongest received power has been taken for granted in the vast literature on cell association. This is for instance the case in the stochastic geometry model of cellular networks [32]. The rationale is clear. This leads to the highest connectivity for each UE to choose the nearest BS unless it is possible to exploit the time-varying fading information, which is often unavailable in practice. Even with the recent emergence of heterogeneous wireless networks, also called *HetNet*, the rule is still valid in terms of coverage probability. That is, a UE is more likely to be covered if it associates with a BS whose received long-term transmission power (called pilot power) is the strongest. A stochastic geometry model to exploit this heterogeneous transmission powers of BSs belonging to multiple tiers in HetNets along with *fading* information has been investigated in [146].

However, from the perspective of load balancing between cells, the rule is invalid in general because each UE might be better off with a lightly-loaded cell rather than heavily-loaded one irrespective of the distances to them. In particular, in HetNet scenarios, it is important to distribute UEs to macro-

cells and micro-cells so that they are equally loaded. The optimal association in the HetNet setting is inherently computationally infeasible, i.e. NP-Hard, whereas the potential gains from load-aware association schemes are much higher [391]. To tackle this problem, a few approximate or heuristic algorithms were proposed based on convex relaxations [391], [381] and non-cooperative and evolutionary games [36], [295]. Most of these algorithms are iterative in nature, requiring many rounds of messaging between UEs and BSs for their convergence.

4.2   Stochastic Network Model

In this chapter, we consider adapting association schemes to optimize any arbitrary function of the SINR received at a single typical UE. To this aim, we first describe a generic stochastic model of the network and define the general performance metric that is induced by an association policy of the UE of interest, which are assumed to be decoupled from those of other UEs. In the rest of the chapter, we use the terminology "technology diversity" to refer to (i) several networks operated on orthogonal bandwidths and (ii) different cellular technologies (e.g. 3G and 4G), both of which can be shared by MVNOs.

Note that extending this framework and results therein to the case where the association policy of a user is affected by those of other users (e.g. load-balancing in HetNet) is mathematically far more challenging and thus is left to future work.

4.2.1 Network Model

We consider T different technologies where T is finite. The BS locations of technology $i \in [1, T]$ are assumed to be a realization of a homogeneous Poisson-Point Process (PPP) ϕ_i on $\mathbb{R}^2$ of intensity λ_i independent of other PPPs. The typical user, from whose perspective we perform the analysis, is assumed to be located at the origin, without loss of generality. Denote by $r_j^i \in \mathbb{R}_+$ the distance to the jth closest point of ϕ_i to the origin, or equivalently the jth nearest BS, where ties are resolved arbitrarily. Hence r_1^i denotes the distance to the closest point (BS) of ϕ_i from the origin.

Each BS of technology i transmits at a fixed power P_i. The received power at a UE from any BS is however affected by fading effects and signal attenuation captured in the propagation model, typically through the path-loss exponent. We assume independent fading, i.e. the collection of fading coefficients H_j^i, which denotes the corresponding value from the jth nearest BS in technology i to the UE, are jointly independent and identically distributed according to some distribution function. We model the propagation path loss through a non-increasing function $l_i(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$, where $i \in \{1, 2, ..., T\}$, i.e. the propagation model for each technology is determined by a possibly different attenuation function. Hence, the signal power received at the typical UE from the jth BS of technology i is $P_i H_j^i l_i(r_j^i)$. For mathematical brevity, we henceforth consider the point process ϕ_i of technology i where each point is *marked* with an independent mark denoting the fading coefficient between the point (BS) to the origin (UE). We can assume that all the random variables

Notation	Brief Description
ϕ_i	Point process corresponding to technology $i \in [1, T]$
λ_i	Intensity (density) of point process ϕ_i
$p_i(\cdot)$	Performance function when associated with technology i
$\mathcal{F}_I$	Information available at the typical UE
j_i	$\arg \sup_{j \geq 1} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})\vert\mathcal{F}_I]$
i^*	The technology chosen by an association policy
$\mathcal{R}_I^\pi$	Average performance of association policy π with $\mathcal{F}_I$
$\mathcal{R}_I^{\pi^*}$	Average performance of optimal policy π^* with $\mathcal{F}_I$

Table 4.1: Table of Notation

belong to a single probability space denoted by $(\Omega, \mathcal{F}, \mathbb{P})$ [44].

4.2.2 Information at a UE

A key concept in this chapter is the tradeoff between the cost of "information" available at UE and the performance gain attained by the association policy making use of that information. Mathematically, the notion of information is encapsulated in a sigma-field $\mathcal{F}_I$ which is a sub-sigma algebra of the sigma-algebra $\mathcal{F}$ on which the marked point processes ϕ_i are defined. A sub-sigma algebra $\mathcal{F}'$ of $\mathcal{F}$ is such that $\mathcal{F}' \subseteq \mathcal{F}$. An example of information is $\mathcal{F}_I = \sigma \left(\cup_{i=1}^{T} \phi_i(B(0, w)) \right)$, which corresponds to the sigma-field generated by the point process up to distance w from the origin. In other words, the UE can estimate BS locations of different technologies r_j^i such that $r_j^i \leq w$. However, note that, we use the information sigma algebra $\mathcal{F}_I$ more generally, which could potentially include fading and shadowing and not just the distances as given in the above example.

4.2.3 Association Policies

An association policy governs the decisions on which technology and BS the typical user (who is located at the origin) should associate with. More formally, an association policy π is a measurable mapping, i.e. $\pi : \Omega \to [1, T] \times \mathbb{N}$ which is $\mathcal{F}_I$ measurable. As stated before, we assume that all additional random variables needed by the policy π are $\mathcal{F}_I$ measurable. The interpretation of the policy π being $\mathcal{F}_I$ measurable is that a typical UE decides to choose a technology and a BS to associate with based only on the information obtainable in the network. It is important to note that while our discussion in this chapter mainly revolves around optimal policies denoted by π^*, our methodology for the performance evaluation in Section 4.5 can be applied for any (suboptimal) policy.

4.2.4 Performance Metrics

All performance metrics considered in this work are functions of SINR (Signal to Interference plus Noise Ratio) received at the typical UE. The SINR of the signal received at the origin from the jth nearest BS of technology i is:

$$\text{SINR}_0^{i,j} = \frac{P_i H_j^i l_i(r_j^i)}{N_0^i + \sum_{k \in \mathbb{N} \setminus \{j\}} P_i H_k^i l_i(r_k^i)},$$

where N_0^i is the thermal noise power which is a fixed constant for each technology $i \in \{1, 2, ..., T\}$. In order to encompass a general set of most useful performance metrics in wireless networks, the performance of different association policies are evaluated through non-decreasing functions of the SINR

observed at the typical UE. Formally, let $p_i(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function for each $i \in \{1, 2, ..., T\}$ which represents the metric of interest if the typical UE associates with technology i. Since π takes values in two coordinates $[1, T] \times \mathbb{N}$ (Section 4.2.3), we divide them into separate coordinates which are denoted by $\pi(0) \in [1, T]$ and $\pi(1) \in \mathbb{N}$, respectively corresponding to the technology and BS chosen by the policy. Then the performance of the association policy π when the information at the typical UE is quantified by $\mathcal{F}_I$ is then given by:

$$\mathcal{R}_I^\pi = \mathbb{E}[p_{\pi(0)}(\mathrm{SINR}_0^\pi)]. \tag{4.1}$$

The subscript I refers to the fact that the information present at the typical UE is $\mathcal{F}_I$. The performance metric $\mathcal{R}_I^\pi$ is averaged over all realizations of the BS deployments, fading variables, and any additional random variables used in the policy π.

Two well-known examples of performance metrics used in practice are coverage probability and average achievable rate. Coverage probability corresponds to setting the function $p_i(x) = \mathbf{1}(x \geq \beta_i)$, which is the chance that the SINR observed at a UE from technology i exceeds a threshold β_i. The other common performance metric of interest, average achievable rate, is defined as $p_i(x) = B_i \log_2(1 + x)$, where the parameter B_i is the bandwidth of technology i. All results on optimal association policy and performance evaluation are stated on the assumption of a general function $p_i(x)$.

4.3 Optimal Association Policy

The optimal association policy denoted by π^* is

$$\pi_I^* = \arg\sup_{\pi} \mathcal{R}_I^\pi, \tag{4.2}$$

where the supremum is over all $\mathcal{F}_I$ measurable policies. From a practical point of view, the optimal association policy is the one that maximizes the performance of the typical UE among all policies having the same "information". In this setup of optimal association, however, we always assume that the typical UE has knowledge of the densities λ_i of the different technologies and the fact that they are independent PPPs although several fundamental results can be easily extended to more general point processes.

Since, we are interested in maximizing an increasing function of the SINR of the typical UE, the optimal association rule is clearly to pick the pair of technology and BS which yields the highest performance conditional on $\mathcal{F}_I$.

Proposition 36. The optimal association algorithm when the information at the typical UE is given by the filtration $\mathcal{F}_I$ is such that

$$\begin{aligned}
\pi_I^*(0) &= \arg\max_{i\in[1,T]} \sup_{j\geq 1} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_I], \\
\pi_I^*(1) &= \arg\sup_{j\geq 1} \max_{i\in[1,T]} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_I],
\end{aligned} \tag{4.3}$$

where the UE must pick the technology $\pi_I^*(0)$ and the $\pi_I^*(1)$-th nearest BS to the origin in $\phi_{\pi_I^*(0)}$.

The performance of the optimal association is

$$\mathcal{R}_I^{\pi^*} = \mathbb{E}[\sup_{j\geq 1} \max_{i\in[1,T]} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_I]]. \tag{4.4}$$

Since $[1, T]$ and $\mathbb{N}$ are countable sets, the order of the maxima in (4.4) does not matter. An important point to observe is that the optimal association given in (4.3) depends on the choice of the performance metric $\{p_i(\cdot)\}_{i=1}^{T}$. Hence, the optimal association rule would be potentially different if one was interested in maximizing coverage probability as opposed to maximizing rate-related metrics for instance.

4.3.1 Ordering of the Performance of the Optimal Association

In this sub-section, we prove an intuitive theorem (Theorem 37) stating that "more" information leads to better performance.

Theorem 37. If $\mathcal{F}_{I_1} \subseteq \mathcal{F}_{I_2}$, then $\mathcal{R}_{I_1}^{\pi^*} \leq \mathcal{R}_{I_2}^{\pi^*}$ where the association rule is the optimal one given in (4.3).

This theorem establishes a partial order on the performance of the optimal policy under different information scenarios at the UE for any performance functions $\{p_i(x)\}_{i=1}^{T}$.

4.3.2 Optimal Association in the Absence of Fading Knowledge

The following lemma is quite intuitive and affirms that the optimal strategy for a UE in the absence of fading knowledge is to associate to the nearest BS of the optimal technology.

Lemma 4.3.1. *If the information $\mathcal{F}_I$ at the typical UE is independent of the sigma algebra generated by the fading random variables, then*

$j_i = \arg\sup_{j \geq 1} \mathbb{E}[p_i(SINR_0^{i,j})|\mathcal{F}_I] = 1$ *and hence* $\pi_I^*(1) = j_{\pi_I^*(0)} = 1$.

4.3.3 Examples of Information

One common class of information is the "*locally estimated information*" which a UE may attain through measurements of (i) received long-term receive pilot signals, which can be easily converted into distances of BS, and (ii) instantaneous received signals, from which fading coefficients can be computed. For example, the knowledge of the distances to BSs no farther than w from the UE is quantified through the sigma-algebra $\mathcal{F}_w = \sigma\left(\cup_{i=1}^{T}\mathcal{F}_w^i\right)$, where $\mathcal{F}_w^i = \sigma\left(\phi_i(B(0,w))\right)$ is the sigma algebra generated by the stochastic process $\phi(B(0,w))$. Furthermore, in case the UE is capable of estimating fading information, one can opt for the sigma-field generated by the marked stochastic process $\phi_i(B(0,w))$, denoted as $\mathcal{F}_w^{i,H}$, where each point (BS) is *marked* with a fading coefficient between the BS and the UE. Here the superscript H refers to the sigma-field generated by the marked point-process.

In existing networks, the most practical example is the knowledge of the nearest L BSs of each technology, denoted by $\mathbf{r}_i^L = [r_i^1, .., r^L]$, i.e. the L-dimensional vector of the distances. In terms of sigma-algebra, it can be defined as $\mathcal{F}_L = \sigma\left(\cup_{i=1}^{T}\mathcal{F}_L^i\right)$, where $\mathcal{F}_L^i = \sigma\left(\phi_i(B(0,r_i^L))\right)$ is the knowledge of the L nearest BS of each technology. One particularly intriguing scenario is complete information about the BS deployments, i.e. $L = \infty$. Denote by $\mathcal{F}_\infty$ the sigma-field for this information scenario and $\mathcal{R}_\infty^{\pi^*}$ as the performance obtained by the optimal policy knowing the entire network. Since $\mathcal{F}_\infty$ is the maximal element among all sub-sigma algebras of $\mathcal{F}$, it follows from Theorem 37 that $\mathcal{R}_\infty^{\pi^*}$ is the the *upper bound* of all achievable performances. To strike

93

a balance between the performance of interest and estimation cost at UE, each MVNO can evaluate $\mathcal{R}_L^{\pi^*}$ to see how much the association policy with L distances stack up against the upper bound $\mathcal{R}_\infty^{\pi^*}$.

4.4   Max-Ratio Association Policy

While the parametric framework in Section 4.3 paves the way for designing the association policy maximizing various metrics, the optimal schemes encapsulated in (4.2) and (4.3) are amenable to tractable analysis only with the knowledge about the underlying PPPs ϕ_i, i.e. their intensities λ_i. On the other hand, it is less conventional at the present time, if not unrealistic, to assume that the densities λ_i are available at the UE in a real network. More importantly, in certain deployment scenarios, it is highly likely that the BS distribution follows a non-homogeneous point process with density (intensity) varying with the location over the network, thereby invalidating the homogeneous PPP assumption.

From the computational perspective, the optimal association can often demand substantial processing power of the UE particularly when the resulting association tailored for a specified performance metric is not simplified into a tractable closed-form expression. In this light, it is desirable to have policies that are completely oblivious to any statistical modeling assumption on the network, i.e. minimalistic policies exploiting universally available information such as distances to BSs, which can be computed from received pilot signal powers in 3G and 4G networks. To address these issues, we propose a *max-*

ratio association policy. This policy has access to the ratio r_2^i/r_1^i information for each technology i, i.e. the information $\mathcal{F}_I = \sigma\left(\cup_{i=1}^T r_2^i/r_1^i\right)$. The max-ratio association is formally described by

$$i^* = \max_{i \in [1,T]} r_2^i/r_1^i, \ j^* = 1.$$

This ratio maximization implies that we place a high priority on a technology where simultaneously the distance to the nearest BS r_1^i is smaller and that to the second-nearest BS r_2^i is larger than other technologies. Note also that the above expression can be easily rearranged into the ratio of the received pilot powers of the nearest and second-nearest BSs when the BS transmission powers within each technology is the same. We show in Theorem 38 that although this policy *per se* is a suboptimal heuristic, it is optimal (in the sense of (4.3)) under a certain limiting regime of the wireless environment.

Theorem 38. Let the noise powers $N_0^i = 0$ for all technologies i and the performance function for all technologies $p_i(\cdot) = p(\cdot)$ for all i. Consider the family of **power-law** path-loss functions $\{l^{(\alpha)}(\cdot)\}_{\alpha>2}$ where $l^{(\alpha)}(x) = x^{-\alpha}$. Let k be **any** integer greater than or equal to 2. If the information at the UE is the k-tuple of the nearest distances of each technology i i.e. $\mathcal{F}_I = \sigma(\cup_{i=1}^T (r_1^i \cdots, r_k^i))$, then

$$\pi_\alpha^*(0) \xrightarrow{\alpha \to \infty} \arg\max_{i \in [1,T]} \frac{r_2^i}{r_1^i} \text{ a.s.,} \qquad (4.5)$$

where π_α^* is the optimal association as stated in (4.3). Recall $\pi_\alpha^*(1) = 1, \forall \alpha$ from Lemma 4.3.1.

This theorem states that max-ratio association is optimal in cases where the signal is drastically attenuated (i.e. large path-loss exponents) with distance, e.g., *metropolitan* or *indoor* environments where the exponent reach values higher than 4, e.g. $\alpha \in [4, 7]$. It is noteworthy that α at higher frequencies as in LTE networks tends to be higher (See, e.g. [179, Chapter 2.6] and references therein). In addition, another implication of this theorem is that it suffices for the asymptotic optimality to exploit the *reduced* information r_1^i/r_2^i per technology in lieu of the given original information, i.e. r_1^i and r_2^i.

4.5 Framework for Performance Analysis

In Section 4.3, we compared the performance of the optimal association policy under different information scenarios by establishing a partial order on them without explicitly computing the performance $\mathcal{R}^\pi$. However, in order to quantify its impact on $\mathcal{R}^\pi$ without resorting to exhaustive simulations, one is also interested in its explicit expression for a given policy π, which may be an optimal policy as in Section 4.3 or a suboptimal one as in Section 4.4. We demonstrate how to explicitly compute $\mathcal{R}^\pi$ in an *automatic* fashion (in Theorem 39) for any arbitrary policy π belonging to a large class of polices, called generalized association, which constitutes another part of our contribution.

4.5.1 Generalized Association

In the rest of the chapter, we restrict our discussion to a class of association policies π that are optimized over information with a special structure

$\mathcal{F}_I$, incorporating what is conventionally available in cellular networks, i.e., we assume that the form of information that a UE has about each technology i is a vector $\mathbf{r}_i \in \mathbb{R}^L$. For instance, if the mobile is informed of the smallest two distances of each technology and their instantaneous signal powers, then $\mathbf{r}_i$ is a 4-dimensional vector with 2 dimensions representing the distances and the other 2 dimensions corresponding to the instantaneous fading powers. That is, we adopt this reduced notation as a surrogate for the sigma-algebra notation in Section 4.3 for simplicity of the exposition. Formally, we assume that the association policy $\pi = \{\pi_i(\cdot)\}_{i=1}^{T}$, according to which a mobile chooses a technology to associate with is given by

$$i^* = \arg \max_{i \in [1,T]} \pi_i(\mathbf{r}_i, \lambda_i), \qquad (4.6)$$

where j_i, the index of BS of technology i to which the UE associates conditioned on selection of i, i.e. $i = i^*$, and $\mathbf{r}_i$ is the L-dimensional vector of observation for technology i.

It is noteworthy that when the technologies are operated on overlapping bandwidths, the above form of association may be extended to a more general form $\pi_i(\{\mathbf{r}_i\}_{i=1}^{T}, \{\lambda_i\}_{i=1}^{T})$, where each association policy utilizes not only the information regarding technology i but also that about all other technologies. Envisioning this extension is easily justifiable because the *desirability* of technology i (represented by $\pi_i(\cdot)$) is affected by the interference inflicted by other technologies. However, we leave it as future work and focus our discuss onto the restricted class of information $\mathcal{F}_I$ in (4.6) which covers the most interesting scenarios of non-overlapping bandwidths.

4.5.2 Performance Computation of the Generalized Association

For each technology i, we denote by $f_i(\mathbf{r}_i)$ the probability density function (pdf) of the information vector $\mathbf{r}_i$ of technology i. For instance, if $L = 1$ and each mobile has knowledge about the location of the nearest base-station r_1^i, then it follows from the property of a PPP that $f_i(\mathbf{r}_i) = f_i(r_1^i)$ is the Rayleigh distribution with parameter $1/\sqrt{2\pi\lambda_i}$. As for the max-ratio policy, $f_i(\mathbf{r}_i) = f_i([r_1^i, r_2^i])$ becomes the distribution of the nearest and second-nearest BSs characterized by the underlying PPP of technology i. We also denote by $f_i^*(\mathbf{r})$ the pdf of the vector $\mathbf{r}_i$ *conditioned* on the event that technology i is selected, i.e. $i^* = i$.

Denote by $f_{\pi_i}(\cdot)$ the pdf of $\pi_i(\mathbf{r}_i, \lambda_i)$ and by $F_{\pi_i}(\cdot)$ the cumulative density function (cdf) of $\pi_i(\mathbf{r}_i, \lambda_i)$. To put it simply, $f_{\pi_i}(\cdot)$ is the cdf of a *function* $\pi_i(\cdot)$ of the given information $\mathbf{r}_i$ rather than that of $\mathbf{r}_i$ itself. For example, in case of the max-ratio policy, $f_{\pi_i}(\cdot)$ is the distribution of r_2^i/r_1^i. To prove the main theorem, we first need to delineate the interplay between the distribution of *optimal* technology $f_i^*(\cdot)$, its original distribution $f_i(\cdot)$, and the (cumulative) distribution of the association policy $F_{\pi_i}(\cdot)$. We have the following lemma from a direct application of Bayes' rule and independence of the point processes ϕ_i.

Lemma 4.5.1. *The probability density function $f_i^*(\mathbf{r})$ is given by*

$$f_i^*(\mathbf{r}) = f_i(\mathbf{r}) \cdot \frac{1}{\mathbb{P}[i^* = i]} \cdot \prod_{j=1, j\neq i}^{T} F_{\pi_j}(\pi_i(\mathbf{r}, \lambda_i)). \tag{4.7}$$

The following theorem finally presents a direct method for computing the performance of any generalized association policy π. Recall that the performance of a policy π is given by $\mathcal{R}_I^\pi = \mathbb{E}[p_{i^*}(\text{SINR}_0^{i^*,j_{i^*}})]$.

Theorem 39. The performance of the association algorithm π under information $\mathcal{F}_I$ denoted by $\mathcal{R}_I^\pi$ is given by-

$$\sum_{i=1}^{T} \int_{\mathbf{r} \in \mathbb{R}^L} \mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}] f_i(\mathbf{r}) \prod_{j=1,j\neq i}^{T} F_{\pi_j}(\pi_i(\mathbf{r}, \lambda_i)) d\mathbf{r}, \qquad (4.8)$$

where $\mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}]$ corresponds to the performance obtained by associating to technology i *conditioned* on the information about technology i the UE has is the vector $\mathbf{r}$.

This theorem states that we need only two expressions, **information** distribution $f_i(\cdot)$ and **policy** distribution $f_{\pi_i}(\cdot)$, in order to derive the performance metric. As exemplified earlier, while $f_i(\cdot)$ is usually a simplistic expression thanks to properties of PPP, mathematical manipulability of $f_{\pi_i}(\cdot)$ highly relies on the complexity of the association policy.

4.6 Computational Examples

In this section, we leverage the results in Section 4.5 to derive several performance metrics in *selected* practical scenarios where the association policy utilizes information $\mathcal{F}_I$ restricted to a vector of distances to BSs ($\mathbf{r}_i$) and BS densities (λ_i) as shown in (4.6). Note however that one can directly compute the performance ($\mathcal{R}_I^\pi$ in Theorem 39) with the probability density function

99

of any association policy ($f_i(\mathbf{r})$) by exploiting Lemma 4.5.1. We show that the resulting performance expressions are mathematically tractable and lend themselves to quantifying the performance of large-scale wireless networks.

For the rest of this section, we consider two representative metrics: (i) *coverage probability* $p_i(x) = \mathbf{1}(x \geq \beta_i)$ and (ii) *average achievable rate* where, to simplify the exposition, we assume the bandwidths of different technologies are the same, i.e. $p_i(x) = p(x) = \log_2(1 + x)$. However, Theorem 39 can be used to compute the performance of any arbitrary non-decreasing function $p_i(\cdot)$. We also assume that the fading variable H_j^i is exponential, i.e. Rayleigh fading, with mean μ^{-1} and the path-loss function $l_i(r) = r^{-\alpha}$ for $\alpha > 2$ for all $i \in [1, T]$. These assumptions have often been adopted for analysis of wireless systems [179] and espoused in stochastic geometry models [44454445].

Let us denote by $c_p(j; \mathbf{r}, \lambda, P, \beta)$ the coverage probability of a UE at the origin served by the jth nearest BS to the origin where the BSs are spatially distributed as a PPP of intensity λ. Here each BS transmits at power P and we are interested in the probabilistic event that the received SINR exceeds the threshold β. The vector $\mathbf{r}$ denotes the vector of distances to BSs, based on which the association decision will be made. More formally,

$$c_p(j; \mathbf{r}, \lambda, P, N_0, \beta) = \mathbb{E}\left[\mathbf{1}\left(\frac{PH_j}{N_0 + \sum_{k \neq j} PH_k} \geq \beta\right)\Big|\mathbf{r}\right] \qquad (4.9)$$

Likewise, we denote by $r(j; \mathbf{r}, \lambda, P)$ the expected rate received by a typical UE at the origin when it is being served by the jth nearest BS to the origin where the BSs are distributed as a PPP of intensity λ and transmitting at power

level P:

$$r(j; \mathbf{r}, \lambda, P, N_0) = \mathbb{E}\left[\log_2\left(1 + \frac{PH_j}{N_0 + \sum_{k \neq j} PH_k}\right)\Big| \mathbf{r}\right]$$
$$= \int_{t \geq 0} c_p(j; \mathbf{r}, \lambda, P, N_0, 2^t - 1)dt. \tag{4.10}$$

Therefore, once we derive an expression for the coverage probability, the average achievable rate expression follows immediately from the calculation of the simple integral in (4.10). In the sequel, we first compute *technology-wise* expressions, (4.9) and (4.10), which are in turn plugged as $p_i(\cdot)$ into Theorem 39 to yield coverage probability metric $\mathcal{R}^{cp}$ and average achievable rate metric $\mathcal{R}^r$, respectively.

4.6.1 Optimal Association Policy

Recall that in the absence of knowledge of fading information, the optimal association policy is to choose technology i^* such that:

$$i^* = \arg\max_{i \in [1,T]} c_p(1; r_i, \lambda_i, P_i, N_0^i, \beta_i), \tag{4.11}$$

$$i^* = \arg\max_{i \in [1,T]} r(1; r_i, \lambda_i, N_0^i, P_i), \tag{4.12}$$

respectively for coverage probability and average achievable rate. Note also that it follows from Lemma 4.3.1 that it is unconditionally optimal to choose the nearest BS for each technology, i.e. $j^* = 1$. Thus our discussion in this section is focused on the choice of technology i.

In this example, we investigate two cases where the UE has knowledge of the distances to the nearest r_1^i or up to the second-nearest BSs $[r_1^i, r_2^i]$ along

with the densities of technologies λ_i while being oblivious to the information about fading H_j^i. In comparison with the standard rule to associate with the nearest BS, this example demonstrates how our proposed framework can be used not only to design an optimal association algorithm maximizing a performance index but also to compute the resulting performance improvements arising from the additional knowledge of distances and densities. The following theorem delineates, among all technologies $i \in [1, T]$, which technology yields the best coverage probability metric.

Theorem 40. If the UE has the knowledge about r_1^i, for all $i \in [1, T]$, the association rule (4.11) with the following expression maximizes the coverage probability:

$$c_p(1; r_1, \lambda, P, N_0, \beta) = \mathrm{e}^{-\mu\beta N_0 r_1^\alpha P^{-1}}$$
$$\exp\left(-2\pi\lambda \int_{u=r_1}^{\infty} \frac{1}{1 + \beta^{-1}(u/r_1)^\alpha} u\, du\right). \quad (4.13)$$

If the UE has the knowledge about r_1^i and r_2^i, for all $i \in [1, T]$, the association rule (4.11) with the following expression maximizes the coverage probability:

$$c_p\left(1; [r_1, r_2], \lambda, P, N_0, \beta\right) = \mathrm{e}^{-\mu\beta N_0 r_1^\alpha P^{-1}} \frac{1}{1 + \beta (r_1/r_2)^\alpha}$$
$$\exp\left(-2\pi\lambda \int_{u=r_2}^{\infty} \frac{1}{1 + \beta^{-1}(u/r_1)^\alpha} u\, du\right). \quad (4.14)$$

To better understand the practical implications of (4.13), we can consider the case where the thermal noise and threshold terms are identical, i.e.

$N_0^i = N_0$ and $\beta_i = \beta$. Since both the first and second factors in the right-hand side of (4.13) are decreasing functions with respect to r_1, the above policy gives preference to smaller r_1^i among all technologies $i \in [1, T]$, which is in line with our intuition.

However, for approximately similar values of r_1^i, it also reveals that the optimal policy tends to choose technology i with lower density λ_i because the right-hand side of (4.13) decreases with λ. The observation is in best agreement with our intuition again because technology i with high density λ_i implies that there are more interfering BSs on the average. On the other hand, the standard rule leads to higher chance of association with the technology with large λ_i because the nearest BS is more likely to belong to the technologies consisting of higher number of BSs. Thus it can be deduced that in case of heterogeneous BS densities, the standard rule leads to very poor coverage performance because of its tendency to associate with the most *populous* technology, whereas the above equation reveals the optimality of associating with *sparsely populated* technology, which sheds light on the complex optimization to be carried out by MVNOs. Likewise, the optimal policy exploiting the additional information of r_2^i exhibits similar tendencies in (4.14) by preferring a technology i having large r_2^i, i.e. the technology with smallest dominant interfering power.

In order to compute the optimal performance metric $\mathcal{R}^{cp}$ resulting from the association rule maximizing the coverage probability, we first need to derive the probability distribution of $c_p(1; r, \lambda_i, P_i, N_0, \beta)$, which is in turn plugged

103

into Theorem 39. The CDF $F_{\pi_i}(y) = \mathbb{P}[c_p(1; r, \lambda_i, P_i, N_0, \beta) \leq y]$ can be simplified into the following expression by using the fact that the nearest distance r of BSs distributed as a PPP is Rayleigh distributed with parameter $\frac{1}{\sqrt{2\pi\lambda_i}}$.

Lemma 4.6.1. *The CDF $F_{\pi_i}(\cdot)$ is given by*

$$F_{\pi_i}(y) = e^{-\ln\left(\frac{1}{y}\right)\frac{1}{2}\left(\int_{v=1}^{\infty} \frac{1}{1+\beta^{-1}(v)^\alpha} v dv\right)^{-1}}. \tag{4.15}$$

Finally, plugging (4.15) into (4.8), we get the following theorem on the coverage probability maximized by the optimal association policy.

Corollary 41. The coverage probability resulting from the optimal association tion exploiting the knowledge of r_1^i is

$$\mathcal{R}^{cp} = \sum_{i=1}^{T} \int_{\mathbf{r} \in \mathbf{R}} c_p(1; r, \lambda_i, P_i, N_0^i, \beta_i) 2\pi\lambda_i r e^{-\pi\lambda_i r^2}$$

$$\prod_{j=1, j\neq i}^{T} \left(\frac{1}{c_p(1; r, \lambda_j, P_j, N_0^j, \beta_j)}\right)^{-\frac{1}{2}\left(\int_{v=1}^{\infty} \frac{1}{1+(\beta_j)^{-1}(v)^\alpha} v dv\right)^{-1}}.$$

4.6.2 Max-Ratio Association Policy

Recall that in the absence of fading information, the Max-Ratio algorithm described in Section 4.4 is to choose technology such that $i^* = \max_{i \in [1,T]} \frac{r_2^i}{r_1^i}$ with the nearest BS in the chosen technology, i.e. $j^* = 1$. Although we saw in

Section 4.6.1 that the density information play a crucial role in performing optimal association, we know from Theorem 38 that the simple non-parametric policy of Max-Ratio is optimal in the limit of large path-loss. In this section, we also show that this policy is tractable and yields expressions for key performance metrics (Corollary 42 and 43). The simplistic form of the policy distribution $F_{\pi_i}(\cdot)$ in the following lemma alludes to ensuing tractable results in this section.

Lemma 4.6.2. *The law $F_{\pi_i}(\cdot)$ for the max-ratio algorithm is:*

$$F_{\pi_i}(x) = \mathbb{P}\left[r_2^i / r_1^i \leq x\right] = 1 - 1/x^2. \tag{4.16}$$

Corollary 42. The coverage probability performance $\mathcal{R}^{cp}$ of the max-ratio algorithm is given by

$$2\sum_{i=1}^{T} \int_{t\geq 1} c_p\left(1; \frac{r_2^i}{r_1^i} = t, \lambda_i, P_i, N_0^i, \beta_i\right) \frac{1}{t^3}\left(1 - \frac{1}{t^2}\right)^{T-1} dt, \tag{4.17}$$

where

$$c_p\left(1; \frac{r_2^i}{r_1^i} = t, \lambda_i, P_i, N_0^i, \beta_i\right) =$$
$$\int_{u=0}^{\infty} c_p\left(1; [u, ut], \lambda_i, P_i, N_0^i, \beta_i\right) 2(\pi\lambda_i)^2 u^3 t^4 e^{-\lambda\pi(ut)^2} du, \tag{4.18}$$

where $c_p\left(1; [u, ut], \lambda_i, P_i, N_0^i, \beta_i\right)$ is given in (4.14).

Since the max-ratio does not optimize a particular performance metric but merely compares the ratio r_2^i / r_1^i, the average achievable rate expression

can be obtained directly from the integral transform in (4.10), which in turn is plugged into Theorem 38 to yield the following corollary.

Corollary 43. The average achievable rate of the max-ratio algorithm is

$$\mathcal{R}^r = 2\sum_{i=1}^{T}\int_{v\geq 0}\int_{t\geq 1} c_p\left(1; \frac{r_2^i}{r_1^i} = t, \lambda_i, P_i, 2^v - 1\right)$$

$$\frac{1}{t^3}\left(1 - \frac{1}{t^2}\right)^{T-1} dtdv, \quad (4.19)$$

where $c_p\left(1; \frac{r_2^i}{r_1^i} = t, \lambda_i, P_i, 2^v - 1\right)$ is given in (4.18).

To get more intuition about the formula, we present the following theorem.

Theorem 44. In the Interference-limited regime (i.e. $N_0^i = 0$ for all $i \in [1, T]$), if the path-loss function is given by $l_i(r) = r^{-\alpha}$ for some $\alpha > 2$ and all $i \in [1, T]$, the coverage probability and the average achievable rate of the max-ratio algorithm are respectively given by

$$\mathcal{R}^{cp} = \begin{cases} \sum_{i=1}^{T}\int_{x=0}^{1} \frac{2(T-1)\cdot x^3(1-x^2)^{T-2}}{1+\beta_i^{2/\alpha}\phi(\alpha,\beta_i,x)}dx, & T \geq 2 \\ \frac{1}{1+\beta_1^{2/\alpha}\phi(\alpha,\beta_1,1)}, & T = 1 \end{cases}, \quad (4.20)$$

$$\mathcal{R}^{r} = \begin{cases} \sum_{i=1}^{T}\int_{x=0}^{1}\int_{t\geq 0} \frac{2(T-1)\cdot x^3(1-x^2)^{T-2}}{1+(2^t-1)^{2/\alpha}\phi(\alpha,2^t-1,x)}dtdx, & T \geq 2 \\ \int_{t\geq 0} \frac{1}{1+(2^t-1)^{2/\alpha}\phi(\alpha,2^t-1,1)}dt, & T = 1, \end{cases} \quad (4.21)$$

where the function ϕ is given by

$$\phi(\alpha, y, x) = \int_{u\geq y^{-2/\alpha}} \frac{1}{1+x^{-\alpha}u^{\alpha/2}}du.$$

106

Since the case $T = 1$ in the above theorem corresponds to the standard rule to associate with the nearest BS in the presence of only one technology, Equations (4.20) and (4.21) reduce to much simpler expressions compared to those in the literature, e.g. Sections III-D and IV-C in the work [32]. On the other hand, as T becomes larger, inside integrand in (4.20), the distribution $2T(T-1)x^3(1-x^2)^{T-2}$ (additional T cancels out the summation operation) is gradually skewed toward the origin $x = 0$, around which $\phi(\alpha, \beta_i, x)$ approaches 0. It is easy to show that the coverage probability $\mathcal{R}^{cp}$ approaches one with higher *technology diversity*, i.e. $T \to \infty$. Though it is not realistic to envision such a large number of technologies or operators, from which each UE can cherry-pick its optimal BS, this theorem demonstrates how much UEs can potentially benefit from the this diversity pooled by MVNOs.

Contrary to the standard association which tends to pick more populous technologies (i.e. large λ_i), giving rise to higher number of interferers, the max-ratio policy counterbalances this pathological behavior by ensuring that the strongest interferer r_2^i is located relatively further. At the same time, the overall performance of max-ratio algorithm critically relies on large path-loss constant α, whereas with this *caveat*, Theorem 38 states that the algorithm is *asymptotically optimal* as $\alpha \to \infty$ for any increasing performance function $p_i(\cdot) = p(\cdot)$ in the interference-limited regime.
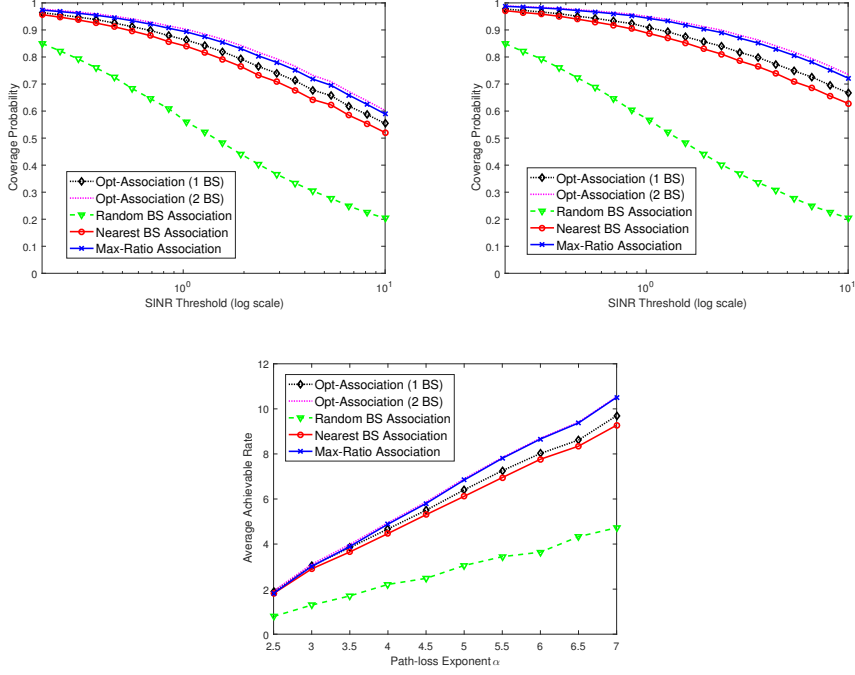
Figure 4.2: Comparison of various association schemes. The first two graphs on the left compare coverage probability where $l_i(r) = r^{-4}$ and $T = 5$ for the first figure and $T = 8$ for the second figure. The rightmost graph compares average achievable rate where α is varied on the x-axis and $l_i(r) = r^{-\alpha}$.

4.7 Simulations and Numerical Results

In this section, we provide more insights into our framework and results by performing simulations and noticing their trends. In performing the simulations, we take as performance metrics, the coverage probability with $p_i(x) = \mathbf{1}(x \geq \beta_i)$ and the average rate with $p_i(x) = \log_2(1 + x)$.

4.7.1 Comparison of Schemes and Technology Diversity

The first two graphs in Fig. 4.2 compare the coverage probability of various association schemes with path-loss exponent $\alpha = 4$ for different number of technologies, $T = 5$ and $T = 8$. We observe in all graphs that the Max-Ratio association scheme outperforms the optimal association policy under the case when only the nearest BS distances are known. More importantly, the Max-ratio association performs *almost as well* as the optimal association under the knowledge of nearest 2 BSs per technology for this typical value of path-loss exponent, not to mention that it outperforms the nearest BS association significantly, particularly when the technology diversity is higher, i.e. $T = 8$.

The rightmost graph in Fig. 4.2 depicts the average achievable rate for path-loss exponents $\alpha \in [2.5, 7]$, which empirically corroborates the statement of Theorem 38 that Max-Ratio is the optimal policy when nearest $k \geq 2$ BS per technology are known in the high path-loss regime. Remarkably, Max-Ratio and the optimal association with two nearest BS distances performs almost equally (indistinguishable in the graph) for $\alpha \geq 5$. That is, a simple non-parametric policy like the max-ratio performs as well as the optimal association policy in which the entire network topology is known (the best possible performance) even in the finite path-loss case. It is also noted that the random BS association, which is the only policy oblivious to technology diversity, results in poor performance in all cases. Thus it is beneficial for MVNOs to leverage the technology diversity in any possible manner by all means.

4.8 Conclusions

In this work, we explored the potential to boost the service performance of wireless networks without incurring additional infrastructure cost by capitalizing on a new form of diversity, which can be either several networks operated on orthogonal bandwidths or multiple wireless technologies pooled by some mobile virtual network operators. We proposed a generic stochastic geometry model for designing association policies that can optimize any desired performance metric. We characterized the optimal policies by establishing a natural monotonicity with respect to increasing information, thereby giving a partial order on the performance without explicit computations. We also proposed a pragmatic data-dependent association policy and showed it to be optimal under reasonable assymptotics. In simulations, we see that the practically reasonable finite network parameters mimic well the limit in our theoretical result, thereby giving great validity to our heuristic association scheme. We believe that our heuristic scheme can serve as an alternative to the standard rules in urban or metropolitan environments with severe signal attenuation which better exploits the new form of diversity.

Chapter 5

Community Detection on Euclidean Random Graphs

5.1 Introduction

In this chapter, we study the problem of graph clustering on euclidean random graphs, from a stochastic geometric perspective and establish fundamental limits and a new clustering algorithm. Community Detection, also known as the graph clustering problem, is the task of grouping together nodes of a graph into representative clusters. This problem has several incarnations that have proven to be useful in various applications ([162]) such as social sciences ([198],[294]), image segmentation [354], recommendation systems ([244],[330]), web-page sorting [221], and biology ([344], [114]) to name a few. In the present thesis, we introduce a new class of spatial random graphs with communities and consider the Community Detection problem on it. Our motivation for a new class of random graph model comes from applications where nodes have geometric attributes, such as in social networks or more generally in graphs of similarities, where the similarity function has metric properties. We study two *regimes* of the random graph - the sparse degree regime and the

---

Parts of this chapter is published in [337] and under review [17]. The author was part of formulating, executing and writing up the results in both papers.

logarithmic degree regime. The sparse degree regime is one wherein the average node degree does not scale with the total number of nodes of the graph, while the logarithmic degree regime is one where the average node degree is proportional to the logarithm of the total number of nodes.

The random graph will be denoted by G_n, which has a random N_n number of nodes which is Poisson distributed with mean λn. In our formulation, $\lambda > 0$ is a fixed constant that denotes the *intensity* parameter and n is a scaling parameter, and we will consider the asymptotic as $n \to \infty$. Nodes are equipped with two i.i.d. labels, a uniform $\{-1, +1\}$ valued *community label* and a uniform $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$, $d \in \mathbb{N}$ valued *location label*. Therefore, the average number of nodes having location labels in any subset of B_n of unit volume is λ, which explains why we call λ as the intensity parameter. To draw the edges, we consider two sequences of functions $(f_{in}^{(n)}(\cdot))_{n \in \mathbb{N}}$ and $(f_{out}^{(n)}(\cdot))_{n \in \mathbb{N}}$ such that $1 \geq f_{in}^{(n)}(r) \geq f_{out}^{(n)}(r) \geq 0$ for all $r \geq 0$ and $n \in \mathbb{N}$. Conditional on the node labels, two nodes with location labels $x, y \in B_n$ and community labels $Z_x, Z_y \in \{-1, +1\}$ are connected by an edge in G_n independently of other edges with probability $f_{in}^{(n)}(||x - y||)$ if $Z_x = Z_y$ and with probability $f_{out}^{(n)}(||x - y||)$ if $Z_x \neq Z_y$. In this interpretation, $|| \cdot ||$ denotes the Euclidean norm on the set B_n if G_n is sparse, or denotes the toroidal metric on B_n in the non-sparse case. We call the random graph G_n *sparse* if the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ do not depend on n. More precisely, the graph G_n is sparse if for all $r \geq 0$, $f_{in}^{(n)}(r) := f_{in}(r)$ and $f_{out}^{(n)}(r) := f_{out}(r)$, for functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying $0 < \int_{x \in \mathbb{R}^d}(f_{in}(||x||) + f_{out}(||x||))dx < \infty$. In

112

this sparse regime, the average degree of any node in G_n is bounded above by $(\lambda/2) \int_{x \in \mathbb{R}^d} (f_{in}(||x||) + f_{out}(||x||))dx < \infty$ *uniformly* in n. In this regime, we draw an edge between two nodes i and j with probability $f_{in}(||X_i - X_j||)$ if the community labels Z_i and Z_j are the same or with probability $f_{out}(||X_i - X_j||)$ if the community labels Z_i and Z_j are different, where $||\cdot||$ denotes the Euclidean norm on the set B_n. Furthermore, the 'boundary effects' due to the edges of the set B_n will not matter asymptotically as $n \to \infty$ as the average degree is uniformly bounded in n (We make this precise in Section 5.2). We call this the sparse regime as the average degree is a constant independent of n. If the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ depended on n and further-more satisfy $\int_{x \in \mathbb{R}^d} f_{in}^{(n)}(||x||)dx = C_{in} \log(n)$ and $\int_{x \in \mathbb{R}^d} f_{out}^{(n)}(||x||)dx = C_{out} \log(n)$ for some $C_{in} > C_{out} \geq 0$, for all n, then we call the graph G_n as the *logarithmic degree* regime or simply as the logarithmic regime. This is so since the average degree of any node is proportional to the logarithm of the total number of nodes in G_n. In this case of logarithmic regime, we avoid having to deal with boundary effects by considering the toroidal metric on the set B_n. Precisely, conditional on the location and community labels on nodes, we place an edge between nodes i and j in G_n with probability $f_{in}^{(n)}(||X_i - X_j||)$ if $Z_i = Z_j$ and with probability $f_{out}^{(n)}(||X_i - X_j||)$ if $Z_i \neq Z_j$, where $||\cdot||$ is the toroidal metric on B_n. We provide a more formal description of the random graph process in both regimes in Section 5.2.

In the sparse regime, we study the problem of 'weak-recovery', which asks how and when one can estimate the community labels of the nodes of

G_n better than at random, with high probability, given observational data of locations labels of all nodes and the graph G_n. Although this requirement on estimation is very weak, we see through our results that this is indeed the best one can hope for in the sparse graph setting considered here. In the logarithmic regime, we consider the problem of 'Strong Recovery' or also known as *exact-recovery*, which asks how and when can one recover the partition of nodes into communities *exactly* based on the observation of the random graph G_n and the location labels of all nodes. As this is a stronger requirement, the graph needs to be *sufficiently dense* in order to perform Exact-Recovery. More precisely, we see that the average node degree must scale logarithmically to the number of nodes to capture the phase transition for exact-recovery. In both of these problems, we assume that the estimator has access to the model parameters λ and $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$. However, we present how one could possibly implement our algorithm in practice, when the connection functions are not known explicitly. From a mathematical perspective, the estimation of the connection functions from data in our spatial setup is an interesting research question which is beyond the scope of this thesis.

5.1.1   Motivations for a New Spatial Graph Model

The most widely studied model for Community Detection is the Stochastic Block Model (SBM), which is a multi-type Erdős-Rényi graph. In the simplest case, the two community symmetric SBM corresponds to a random graph with n nodes, with each node equipped with an i.i.d. uniform community label

drawn from $\{-1, +1\}$. Conditionally on the labels, pairs of nodes are connected by an edge independently of other pairs with two different probabilities depending on whether the end points are in the same or different communities. Structurally, the sparse SBM is known to be locally tree-like ([288],[16]) while real social networks are observed to be *transitive* and sparse. Sparsity in social networks can be understood through 'Dunbar's number' [149], which concludes that an average human being can have only about 500 'relationships' (online and offline) at any point of time. Moreover, this is a fundamental cognitive limitation of the person and not that of access or resources, thereby justifying models where the average node degree is independent of the population size. Social networks are transitive in the sense that any two agents that share a mutual common neighbor tend to have an edge among them as well, i.e., the graph has many triangles. Similar phenomena also takes place in graphs of similarities, where vertices are connected based on metric similarity functions. These aspects point out the limitations of the sparse SBM and a large collection of models have been proposed to better fit applications under the realm of Latent Space Models ([191],[197]) and inhomogeneous random graphs ([83]). See also [16] for more references. These are sparse *spatial* graphs in which the agents of the social network are assumed to be embedded in an abstract *social space* that is modeled as an Euclidean space and conditional on the embedding, edges of the graph are drawn independently at random as a non-increasing function of the distance between two nodes. Thanks to the properties of Euclidean geometry, these models are transitive and sparse, and

115

have a better fit to data than any SBM ([191]). Such modeling assumptions in the context of multiple communities was also recently verified in parallel independent work [171], where the nodes have both a community label and a location label. The locations labels are sampled uniformly on a sphere and the edges are generated by nodes 'nearby' in this sphere connecting with probabilities that depend on the community labels. Several empirical validations of this model on real data is also conducted in [171] which suggests that such spatial random graph model provides a good fit for several real world networks. However, we note that the sparse SBM enjoys certain advantages over the geometric random graph considered here, namely that of having low diameter, in agreement with the 'small world' phenomena observed in many real world networks (see [371]). Therefore a natural next step is to superimpose an SBM with the type of geometric graphs considered here to obtain both a lot of triangles and small diameter, i.e. a type of small world SBM.

Thus, one can view our model as the simplest planted-partition version of the Latent Space model, where the nodes are distributed uniformly in a large compact set B_n and conditional on the locations, edges are drawn depending on Euclidean distance through connection functions $f_{in}.(\cdot)$ and $f_{out}(\cdot)$. Although, our assumptions are not particularly tailored towards any real datasets, our setting is the most challenging regime for the estimation problem as the location labels alone without the graph reveal no community membership information. However,in this thesis we assume the location labels on nodes are known exactly to the estimator. In practice, it is likely that the locations

labels are unknown (as in the original Latent Space models where the social space is unobservable) or are at-best estimated separately. Nonetheless, our formulation with known location labels forms a crucial first step towards more general models where the location labels are noisy or missing. The problem with known spatial location labels is itself quite challenging as outlined in the sequel and hence we decided to focus on this setting alone in the present thesis. Another drawback of our formulation is that we assume the estimator has knowledge of the model parameters $f_{in}(\cdot)$ and $f_{out}(\cdot)$. In our spatial setup, the estimation of connection functions from data is an interesting research question in itself which is however beyond the scope of this thesis.

Central Technical Challenges - The core technical challenge in studying our spatial graph model in the sparse regime lies in the fact that it is not 'locally tree-like'. The spatial graph is locally dense (i.e., there are lots of triangles) which arises as a result of the constraints imposed by Euclidean geometry, while it is globally sparse (i.e., the average degree is bounded above by a constant). The sparse SBM on the other hand, is locally 'tree-like' and has very few short cycles [288]. This comes from the fact that the connection probability in a sparse SBM scales as c/n for some $c > 0$. In contrast, the connection function in our model in the sparse regime does not scale with n. From an algorithmic point of view however, most commonly used techniques (message passing, broadcast process on trees, convex relaxations, spectral methods etc) are not straight forward to apply in our setting since their analysis fundamentally relies on the locally tree-like structure of the graph (see [16] and refer-

ences therein). This fundamental difficulty renders the weak-recovery problem in the sparse regime quite challenging, even in the presence of known spatial location labels. We overcome this difficulty by proposing a novel clustering algorithm and lower bound technique. The key idea for our algorithm is to exploit the fact that our graph is locally quite dense, which allows one to classify very accurately 'nearby' pairs of nodes. We then use ideas from percolation theory to then piece together in a non-trivial fashion the different nearby estimators to produce a global clustering. To prove the lower bound, we develop a new coupling argument by connecting Community Detection with a problem we call 'Information Flow from Infinity', which is a new problem and is of independent interest. In the logarithmic degree regime, our main novelty is to notice that the lower bound for community detection can be reduced to reasoning about events in disjoint regions of space. We then exploit this reduction along with the fact that events of a Poisson process in disjoint regions of space are independent, along with certain symmetries of the Euclidean space, to provide an explicit lower bound for exact-recovery. In certain cases, we give a closed form expression of the phase-transition threshold by using the large deviations framework introduced in [20].

The other speciality in our setting is the presence of location labels which are known exactly to the estimator. This provide some form of 'side-information' which any estimator must exploit. As an illustrative example to see this, consider the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ in the sparse case to be of bounded support. In this case, the absence of an edge between two nearby

118

nodes makes it likely that these two nodes belong to opposite communities but the lack of an edge between far-away nodes farther than the support of either connection functions does not give any community membership information. Thus the lack of an edge in this example has different interpretations depending on the location labels which needs to be exploited in a principled manner by the estimator. Indeed this is best seen in our lower bound for exact-recovery case, where if $f_{out}^{(n)}(r)$ is identically 0, i.e. there are no cross community edges, then Exact-Recovery might still be possible even before the subgraph on the nodes of each individual communities become fully connected. This takes place since we can use the spatial location information in a non-trivial fashion to estimate the labels of isolated nodes. Nonetheless, the location labels alone without the graph provide no information on community membership as the community and location labels are independent of each other.

5.1.2 Related Work

Community Detection on sparse graphs has mostly been studied on the SBM random graph model. The study of SBM has a long history in the different literatures of statistical physics (see [142] and references therein), mathematics (for ex. [82],[288]) and computer science (for ex. [265], [128],[286]). The reader should refer to the survey [16] for further background and references on the SBM. The survey [281] gives a complete treatment of the SBM from a statistical physics view point. There has been renewed interest in the sparse regime of the SBM following the paper [142], which made a number of striking

conjectures on phase-transitions. Subsequently, some of them have been established with the most notable and relevant achievements to ours being that of [288],[284], [262] and [91]. These papers prove that both Community Detection and the distinguishability problem for the two community sparse SBM undergo a phase-transition at the same point which they characterize explicitly. These results for the SBM motivates the investigation of the phase-transitions for Community Detection and distinguishability in our model. However, the tools needed for our model are very different from those used to study the SBM. The key ideas for all of our results come from different problems, mostly those studied in the theory of percolation [83] and stochastic geometry [303]. Our algorithm is motivated by certain ideas that appeared in the Interacting Particle Systems literature (for ex. [303] [241],[304],[150]). The papers there developed re-normalization and percolation based ideas to study different particle systems arising in statistical mechanics, and our analysis bears certain similarity to that line of work. Our lower bound comes from identifying an *easier* problem than Community Detection called Information Flow from Infinity, which is a new problem. The key idea to show this reduction comes from an ergodic argument applied on the spatial random graphs. To study the impossibility of Information Flow from Infinity, we employ certain 'random-cluster method' and coupling arguments. Such methods are quite popular and have proven to be extremely fruitful in other contexts for example to study mixing time of Ising Models ([254],[253],[252]). Similar coupling ideas have also appeared in other estimation contexts, notably the reconstruction on trees problem [282],

where a lower bound was established using this method.

In the non-sparse regime, the work of [18] and [20] is most related to our methods for the non-sparse regime. Both papers studies the problem of Exact-Recovery in much greater detail in the generalized SBM, proving both lower bounds and efficient algorithms for exact recovery. In this thesis, we use the lower bound framework developed in [20], specifically, the large deviations characterization of hypothesis testing between Poisson random vectors, to provide a lower bound for our spatial random graph model.

From a modeling perspective, the works of [389] and [171] which considers the Latent Space Models as a part of its model is the closest to our model. The paper of [389] gives a spectral algorithm that is proven to work in the logarithmic degree regime. In particular, their methods do not work for the sparse regime which is one of the central topics discussed in the present thesis. The algorithm provided in [171] bears certain similarity to ours where nearby nodes' community membership are tested based on the common neighbors. However, it is only guaranteed to work in the logarithmic degree regime. Furthermore, no lower bounds are presented in [171], while the lower bound in [389] is based on the idea that the graph is locally tree-like, which is not the case in our model. Another related problem of group synchronization was introduced in [19], which among other things, also considered the question of how well can one identify the communities of two 'far-away' nodes. This paper establishes certain phase-transitions for weak recovery that look similar to our lower bound, using different ideas from percolation and random walks

on grids. Nonetheless, our algorithm is completely different from theirs and in-fact a straight forward adaptation of our algorithm to their setting can give an alternative proof of Theorem 3 in [19].

5.2 Mathematical Framework and Problem Statement

We describe the mathematical framework based on stationary point processes and state the problem of Community Detection. We will consider the graph G_n as a suitable truncation of an infinite random graph constructed on the support of a Poisson Point Process (PPP). This representation allows us to couple the graphs G_n for all n on a single probability space, which aids in the mathematical analysis. In the non-sparse case, there is no direct limiting infinite graph. Nevertheless, we can couple all the graphs G_n on a single probability space by constructing them on a single marked PPP. We set a common shorthand notation we use throughout. For two arbitrary positive sequences $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$, we let $b_n = o(a_n)$ to denote the fact that $\lim_{n \to \infty} b_n / a_n = 0$.

5.2.1 The Planted Partition Random Connection Model

We suppose there exists an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we have an appropriately marked PPP $\bar{\phi}$. We will construct the sequence of random graphs $(G_n)_{n \in \mathbb{N}}$ on this space simultaneously for all n as a measurable function of this marked PPP $\bar{\phi}$. More formally, we assume ϕ to be the support of a homogeneous PPP of intensity λ on $\mathbb{R}^d$ with the enumeration that $\phi :=$

122

$\{X_1, X_2, \cdots\}$. In this representation, $X_i \in \mathbb{R}^d$ for all $i \in \mathbb{N}$. Furthermore, we assume the enumeration to be such that for all $i > j \in \mathbb{N}$, we have $||X_i||_\infty \geq ||X_j||_\infty$, i.e., the points are ordered in accordance to increasing l_∞ distance. We further mark each atom $i \in \mathbb{N}$ of ϕ with random variables $Z_i \in \{-1, +1\}$ and $\{U_{ij}\}_{j \in \mathbb{N}\setminus\{i\}} \in [0, 1]^{\mathbb{N}\setminus\{i\}}$ satisfying $U_{ij} = U_{ji}$ for all $i \neq j \in \mathbb{N}$. We denote by $\bar{\phi}$ to be this marked PPP. The sequence $\{Z_i\}_{i \in \mathbb{N}}$ is i.i.d. with each element being uniformly distributed in $\{-1, +1\}$. For every $i \in \mathbb{N}$, the sequence $\{U_{ij}\}_{j \in \mathbb{N}\setminus\{i\}}$ is i.i.d. with each element of $\{U_{ij}\}_{j \in \mathbb{N}\setminus\{i\}}$ being uniformly distributed on $[0, 1]$. The interpretation of this marked point process is that for any node $i \in \mathbb{N}$, its location label is X_i, community label is Z_i and $\{U_{ij}\}_{j \in \mathbb{N}\setminus\{i\}}$ are used to sample the graph neighbors of node i. We describe this construction in both the sparse and logarithmic degree regimes below.

Denote by the set $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2} \right]^d$, the cube of area n in $\mathbb{R}^d$. For all $n \in \mathbb{N}$, we let $N_n := \sup\{i \geq 1 : X_i \in B_n\}$. Since the nodes are enumerated in increasing order of l_∞ distance, it follows that for all $i \in [1, N_n]$, $X_i \in B_n$. Furthermore, from basic properties of PPP, N_n is a Poisson random variable of mean λn. We construct the graph G_n, by assuming its vertex set to be $\{1, \cdots, N_n\}$ and the location label of any node $i \in [1, N_n]$ to be $X_i \in B_n$ and its community label to be $Z_i \in \{-1, +1\}$. However, we use the marks $\{U_{ij}\}_{j \in \mathbb{N}\setminus\{i\}}$ slightly differently depending on whether the graph is sparse or not. Recall that in the sparse regime, we have the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ to be independent of n. In this case, we first construct an infinite graph G with vertex set $\mathbb{N}$ and place an edge between any two nodes $i, j \in \mathbb{N}$

123

if and only if $U_{ij} = U_{ji} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i \neq Z_j}$. The graph G_n is then the induced subgraph of G consisting of the nodes 1 through N_n, i.e., the subgraph of G restricted to the node set with location labels in B_n. In the logarithmic degree regime, we assume that the set B_n is equipped with the toroidal metric rather than the Euclidean metric for simplicity. Formally, for any $x := (x_1, \cdots, x_d), y := (y_1, \cdots, y_d) \in B_n$, the toroidal distance on B_n is given by $||x - y||_{\mathfrak{I}_n} = ||(\min(|x_1 - y_1|, n^{1/d} - |x_1 - y_1|), \cdots, \min(|x_d - y_d|, n^{1/d} - |x_d - y_d|))||$, where $||.||$ is the standard Euclidean norm on $\mathbb{R}^d$. For any $i, j \in [1, N_n]$, we draw an edge between nodes i and j in G_n if and only if $U_{ij} = U_{ji} \leq f_{in}^{(n)}(||X_i - X_j||_{\mathfrak{I}_n})\mathbf{1}_{Z_i=Z_j} + f_{out}^{(n)}(||X_i - X_j||_{\mathfrak{I}_n})\mathbf{1}_{Z_i \neq Z_j}$.

The infinite random graph G in the sparse regime can be viewed as a 'planted-partition' version of the classical random-connection model ([268]). Given $\lambda \in \mathbb{R}_+$, $g(\cdot) : \mathbb{R}_+ \to [0, 1]$ and $d \geq 1$, the classical random-connection model $H_{\lambda, g(\cdot), d}$, is a random graph whose vertex set forms a homogeneous PPP of intensity λ on $\mathbb{R}^d$. Conditionally on the locations, edges in $H_{\lambda, g(\cdot), d}$ are placed independently of each other where two points at locations x and y of $\mathbb{R}^d$ are connected by an edge in $H_{\lambda, g(\cdot), d}$ with probability $g(||x - y||)$. This construction can be made precise by letting the edge random variables for each node be marks of the PPP similarly to the construction of G.

5.2.2   The Community Detection Problem

In this thesis, we study two different notions of Community Detection - weak recovery and exact recovery, depending on whether the graph G_n is

sparse or non-sparse. In the sparse regime, our definition of Community Detection is analogous to the notion of 'weak-recovery' considered in the classicial SBM literature ([142]) and in the logarithmic degree regime, our definition of Community Detection is analogous to the notion of 'Exact-Recovery' ([18]) in the SBM literature. To state the two notions of community recovery, we set more notation. Let ϕ_n be the restriction of the point process ϕ to the set B_n. Notice that the cardinality of ϕ_n is N_n which is distributed as a Poisson random variable of mean λn, and $X_i \in B_n$ for all $i \in [1, N_n]$. Moreover, conditionally on N_n, the location variables $(X_i)_{i \in [1, N_n]}$ are placed uniformly and independently in B_n. Before describing the problem, we need the definition of 'overlap' between two sequences.

Definition 45. Given a $t \in \mathbb{N}$, and two sequences $\mathbf{a}, \mathbf{b} \in \{-1, 1\}^t$, the *overlap* between $\mathbf{a}$ and $\mathbf{b}$ is defined as $\frac{|\sum_{i=1}^{t} a_i b_i|}{t}$, i.e., the absolute value of the normalized scalar product.

We define two notions of performance of community detection, weak recovery and exact recovery defined below.

Definition 46. *Weak Recovery* is said to be **solvable in the sparse regime** for $\lambda, d, f_{in}(\cdot)$ and $f_{out}(\cdot)$ if for every $n \in \mathbb{R}_+$, there exists a sequence of $\{-1, +1\}$ valued random variables $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ which is a deterministic function of the observed data G_n and ϕ_n such that there exists a constant $\gamma > 0$ satisfying

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{O}_n \geq \gamma\right] = 1, \tag{5.1}$$

125

where $\mathcal{O}_n$ is the overlap between $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ and $\{Z_i\}_{i=1}^{N_n}$.

In the above definition, we let the overlap $\mathcal{O}_n := 1$ if $N_n = 0$. In the logarithmic degree regime, we ask for exact-recovery which is formally stated as follows.

Definition 47. *Exact-Recovery* is said to be **solvable in the logarithmic degree regime** for $\lambda, d, (f_{in}^{(n)}(\cdot))_{n \in \mathbb{N}}$ and $(f_{out}^{(n)}(\cdot))_{n \in \mathbb{N}}$ if for every $n \in \mathbb{R}_+$, there exists a sequence of $\{-1, +1\}$ valued random variables $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ which is a deterministic function of the observed data G_n and ϕ_n such that

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{O}_n = 1\right] = 1, \tag{5.2}$$

where $\mathcal{O}_n$ is the overlap between $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ and $\{Z_i\}_{i=1}^{N_n}$.

A key new feature of our Definitions 46 and 47 comes from our assumption that the algorithm has knowledge of all location labels on the nodes and it only needs to estimate the missing community labels. In the sparse regime, we ask when can any estimator assign community labels to the nodes that beats a 'random guess'. Observe that if an estimator guessed every node to be in Community $+1$, then the achieved overlap $\mathcal{O}_n$ converges almost-surely to 0, thanks to the Strong Law of Large Numbers. Thus, achieving a positive γ asks whether an estimator can asymptotically beat the trivial estimator. In the non-sparse regime, we ask when and how can we recover the commuity label of *all* nodes, upto a global flip.

126

Observe that we assume the algorithm has access to the parameters $f_{in}^{(n)}(\cdot)$, $f_{out}^{(n)}(\cdot)$ and λ although this assumption may not always hold in practice. As mentioned, the estimation of model parameters from data itself will form an interesting technical question which we leave for future work. We take an absolute value in the definition of overlap since the distribution of G_n is symmetric in the community labels. In particular, if we flipped all community labels of G_n, we would observe a graph which is equal in distribution to G_n. Thus, any algorithm can produce the clustering only up-to a global sign flip, which we capture by considering the absolute value. We take finite restrictions B_n since the overlap is not well defined if $N_n = \infty$. A natural question then is of 'boundary-effects', i.e. the nodes near the boundary of B_n will have different statistics for neighbors than those far away from the boundary. However since G_n is sparse, except for a $o_n(1)$ fraction of nodes, all nodes in G_n will have the same degree as in the infinite graph G, i.e. the boundary effects are negligible. In the non-sparse case, since we consider the set B_n as a torus, to precisely avoid the technicalities arising out of considering the edge effects.

The following elementary monotonicity property is evident from the definition of the problem and sets the stage for stating our main results.

Proposition 48. For every $f_{in}(\cdot)$, $f_{out}(\cdot)$ and d, there exists a $\lambda_c \in [0, \infty]$ such that

- $\lambda < \lambda_c \implies$ weak-recovery is not solvable.

- $\lambda > \lambda_c \implies$ there exists an algorithm (which could possibly take exponential time) to solve weak-recovery.

Proof. Assume, for a given value of λ, there exists a community detection algorithm that achieves an overlap of $\gamma > 0$. Now, given any $\lambda' > \lambda$, we will argue that we can achieve positive overlap. The proof of this follows from the basic thinning properties of the PPP. Given an instance of the problem with intensity λ', we will remove every node along with its incident edges independently with probability $1 - \frac{\lambda}{\lambda'}$. We assign a community label estimate of $+1$ to all the removed nodes. For the nodes that remain (which is then an instance of the problem of community detection with intensity λ), we achieve an overlap of γ with probability at-least $1 - o_n(1)$, from the hypothesis that we can achieve positive overlap at intensity λ. Thus, from the independence of the thinning procedure and the community labels and strong law of large numbers, the overlap achieved by this process on an instance of intensity λ' will be at-least $\frac{\lambda \gamma}{\lambda'}$ with probability at-least $1 - o_n(1)$. Thus, the problem of community detection solvability is monotone in λ in the sparse case. □

This proposition is not that strong since it does not rule out the fact that λ_c is either 0 or infinity. Moreover, this proposition does not tell us anything about polynomial time algorithms, just of the existence or non-existence of any (polynomial or exponential time) algorithms. The first non-trivial result would be to establish that $0 < \lambda_c < \infty$, i.e. the phase transition is strictly

non-trivial and then to show that for possibly a larger constant, the problem is solvable efficiently. We establish both of this in Section 5.3.

Distinguishability of the Planted Partition in the Sparse Regime

In this study, we also consider a related problem to weak-recovery, namely the *distinguishability* question in the sparse regime.

Definition 49. The sparse planted partition model with parameters λ, d and connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ is said to be *distinguishable*, if for every $g(\cdot) : \mathbb{R}_+ \to [0, 1]$, one can identify with probability at-least $\frac{1}{2} + \gamma - o_n(1)$ for some $\gamma > 0$, whether the observed data (G_n, ϕ_n) is a sample of the planted partition with the above parameters, or is a sample of the random connection model $H_{\lambda, g(\cdot), d}$, given an uniform prior over the two models.

This problem asks if we can even identify the existence of communities, before trying to identify them. Such hypothesis testing questions are of critical importance in practice where one needs to be reasonably sure of the presence of communities in a given graph, before attempting to cluster the graph. Our main result in Theorem 56 is that one can always solve the distinguishability problem, i.e. it exhibits no phase-transition. Thus our results on weak-recovery and distinguishability predicts regimes of the problem (i.e. $d = 1$ and $\lambda < \lambda_c$ for $d \geq 2$), where we can be very sure of the presence of communities, but cannot recover it better than at random.

129

5.3 Main Results

5.3.1 Lower Bound for Weak Recovery

To state the main lower bound result in Theorem 50 below, we set some notation. For the random connection model graph $H_{\lambda,g(\cdot),d}$, denote by $\mathcal{C}_{H_{\lambda,g(\cdot),d}}(0)$ the set of nodes of $H_{\lambda,g(\cdot),d}$ that are in the same connected component as that of the node at the origin under the measure $\mathbb{P}^0$. Denote by $\theta(H_{\lambda,g(\cdot),d}) := \mathbb{P}^0[|\mathcal{C}_{H_{\lambda,g(\cdot),d}}(0)| = \infty]$ the percolation probability of the random graph $H_{\lambda,g(\cdot),d}$, i.e. the probability (under Palm) that the connected component of the origin has infinite cardinality.

Theorem 50. If $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) = 0$, then weak-recovery is not solvable.

The proof of this Theorem is presented in Section 5.5. This theorem states that if the two functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are not 'sufficiently far-apart', then no algorithm to detect the partition of nodes can beat a random guess. As a corollary, this says that Community Detection is impossible for $d = 1$.

Corollary 51. For all $\lambda > 0$, $f_{in}(\cdot), f_{out}(\cdot)$ such that $\int_{x \in \mathbb{R}} f_{in}(||x||)dx < \infty$, weak-recovery is not solvable if $d = 1$.

Proof. This is based on the classical fact that for all $g(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\int_{x \in \mathbb{R}} g(||x||)dx < \infty$, $\theta(H_{\lambda,g(\cdot),1}) = 0$. $\qquad\square$

The following corollary gives a quantitative estimate of the percolation probability for higher dimensions in terms of the problem parameters.

Corollary 52. For all $d \geq 2$, if $\lambda \leq \lambda_{lower} := (\int_{x \in \mathbb{R}^d} (f_{in}(||x||) - f_{out}(||x||)) dx)^{-1}$, then weak-recovery cannot be solved. Thus, $\lambda_c > (\int_{x \in \mathbb{R}^d} (f_{in}(||x||) - f_{out}(||x||)) dx)^{-1}$.

Proof. From classical results on percolation [268], by comparison with a branching process, we see that $\lambda \int_{x \in \mathbb{R}^d} g(||x||) \leq 1 \implies \theta(H_{\lambda, g(\cdot), d}) = 0$. $\qquad\qquad\square$

Recall that if the graph G is sparse, then $\int_{x \in \mathbb{R}^d} f_{in}(||x||) dx < \infty$ which implies from Corollary 52 that λ_c is strictly positive in the sparse regime. The following proposition shows that this lower bound is tight for certain specific families of connection functions.

Proposition 53. For all $d \geq 2$ and $R_1 \geq R_2$, if $f_{in}(r) = \mathbf{1}_{r \leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_2}$ and λ is such that $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$, then weak-recovery can be solved in time proportional to n with the proportionality constant depending on the parameters λ, R_{in} and R_{out}.

The proof of this Proposition is presented in Section 5.5. Hence, in view of Theorem 50, for $f_{in}(r) = \mathbf{1}_{r \leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_2}$ for some $R_1 \geq R_2$, then if λ is such that $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 0$, no algorithm (exponential or polynomial) time can solve weak-recovery, while if λ is such that $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$, then a linear time algorithm exists to solve weak-recovery. This gives a sharp phase-transition for this particular set of parameters where the problem shifts from being unsolvable even with unbounded computation to being efficiently solvable.

5.3.2   Algorithm and an Upper Bound for Weak Recovery

Our main result in the positive direction is our GBG algorithm described in Section 5.4. The main theorem statement on the performance of GBG is the following.

Theorem 54. If $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are such that $\{r \in \mathbb{R}_+ : f_{in}(r) \neq f_{out}(r)\}$ has positive Lebesgue measure and $d \geq 2$, then there exists a $\lambda_{upper} < \infty$ depending on $f_{in}(\cdot), f_{out}(\cdot)$ and d, such that for all $\lambda \geq \lambda_{upper}$, the GBG algorithm solves the weak-recovery problem. Moreover, GBG when run on data (G_n, ϕ_n), has time complexity order n^2 and storage complexity order n.

The proof of this Theorem is in Appendix D. This theorem gives a complete non-trivial phase-transition for the sparse graph case where we have $0 < \lambda_c < \infty$ which implies the existence of different phases. We also note that our algorithm is asymptotically optimal in a weak sense made precise in the sequel below. Denote by $\mathcal{O}_\lambda$ as the maximum overlap achieved by our algorithm, with the definition of overlap as given in Definition 45. More precisely, denote by $\mathcal{O}_\lambda$ as

$$\mathcal{O}_\lambda := \sup\{\gamma \geq 0 : \lim_{n \to \infty} \mathbb{P}[\mathcal{O}_n > \gamma] = 1\}, \tag{5.3}$$

where $\mathcal{O}_n$ is the overlap achieved by the GBG algorithm when run on the data G_n and ϕ_n. Thus, for each fixed λ, the overlap $\mathcal{O}_\lambda \in [0,1]$ is determinstic quantity.

Proposition 55. The following limit $\lim_{\lambda \to \infty} \mathcal{O}_\lambda = 1$ exists.

The proof is presented in Appendix D. In words, this proposition states that as the 'signal' gets stronger, the fraction of nodes correctly classified by our algorithm tends to 1. On a related note, we also mention in Section 5.4.5, a practical way of implementing the algorithm when the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are not known explicity.

5.3.3 Distinguishability of the Planted Partition

The key result we show here is that unlike in the traditional Erdős-Rényi setting, the planted partition random connection model is always mutually singular with respect to *any* random connection model without communities. Before precisely stating the result, we set some notation. Denote by $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$ the Polish space of all simple spatial graphs whose vertex set forms a locally finite set of $\mathbb{R}^d$. Thus, our random graph G or the random connection model $H_{\lambda, g(\cdot), d}$ can also be viewed through the induced measure on the space $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$.

Theorem 56. For every $\lambda > 0$, $d \in \mathbb{N}$ and connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying $1 \geq f_{in}(r) \geq f_{out}(r) \geq 0$ for all $r \geq 0$, and $\{r \geq 0 : f_{in}(r) \neq f_{out}(r)\}$ having positive Lebesgue measure and $g(\cdot) : \mathbb{R}_+ \to [0, 1]$, the probability measures induced on the space of spatial graphs $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$ by G and $H_{\lambda, g(\cdot), d}$ are mutually singular.

The proof of this Theorem is presented in Appendix D. This theorem 56 implies that this distinguishability problem as stated in Definition 49 can be solved with probability of success $1 - o_n(1)$ for *all* parameter values. Thus,

the distinguishability problem as stated in our spatial case exhibits no phase-transition. A consequence of our results is that in certain regimes ($\lambda < \lambda_c$ for $d \geq 2$ and $\lambda > 0$ for $d = 1$), we can be very sure by observing the data that a partition exists, but cannot identify it better than at random. Such phenomena was proven not to be observed in a symmetric SBM with two communities ([288]) and conjectured not to occur in any arbitrary SBM ([142]). Technically, this theorem gives in particular that G and $H_{\lambda, \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}, d}$ are mutually singular. Note that if $g(\cdot) \neq \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}$, then the average degrees of G and $H_{\lambda, g(\cdot), d}$ are different and hence the empirical average of the degrees in G_n and $H_{\lambda, g(\cdot), d}$ restricted to B_n, will converge almost surely as $n \to \infty$ (thanks to the ergodic property of PPP) to the mean degree, thereby making the two induced measures mutually singular Thus, the only non-trivial random connection model that can possibly be not singular with respect to G is $H_{\lambda, \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}, d}$, i.e. the case of equal average degrees. We show by a slightly different albeit similar ergodic argument in Section D.2 that even in the case of equal average degrees, the two induced measures are mutually singular.

5.3.4 Phase Transition for Exact-Recovery

Our main achievement with regards to Exact-Recovery is an explicit necessary condition on the model parameters given in the following Theorem. To highlight the ideas, we primarily focus on a simplest non-trivial example of connection functions in the logarithmic regime, where we are able to conjecture a closed form expression for the phase-transition threshold.

Definition 57. For any $0 \leq b < a \leq 1$, $n \in \mathbb{N}$, $\lambda > 0$ and $d \in \mathbb{N}$, we denote by $\mathcal{G}(\lambda n, a, b, d)$ as the distribution of graph G_n we defined in Section 5.2 with $f_{in}^{(n)}(r) = a\mathbf{1}_{r \leq \log(n)^{1/d}}$ and $f_{out}^{(n)}(r) = b\mathbf{1}_{r \leq \log(n)^{1/d}}$ for all $r \geq 0$, where the distance on the set B_n is the torridal metric.

In the results that follow, denote by ν_d to be the volume of the unit Euclidean unit ball in d dimensions.

Theorem 58. For any $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is not solvable.

The proof is presented in Appendix D. This connection functions are the simplest non-trivial instance of our model that makes the study interesting. We also believe that this necessary condition to be tight. However, at this point, we only present the following conjecture and do not pursue a proof of this.

Conjecture 59. For any $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is solvable.

We believe two-round techniques developed in [20] applied to the spatial graph case can be fruitful in establishing this conjecture. To establish the existence of a phase-transition however, we analyze the `GBG` algorithm and adapt it to the non-sparse case to yield the following result.

135

Theorem 60. For every $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$, there exists $C(a, b, d) > 0$ such that if $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > C(a, b, d)$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is solvable by `GBG` algorithm.

The proof is presented in Appendix D. This theorem gives the existence of different phases of the exact-recovery problem depending on λ. In particular, it states that if the intensity λ is sufficiently high, then Exact-Recovery is solvable by our `GBG` algorithm.

5.4 Algorithm for Performing Community Detection

In this section, we outline an algorithm called `GBG` described in Algorithm 3 that has time complexity of order n^2 and storage complexity of order n. We make the presentation here assuming that G_n is sparse, although a straightforward adaptation can be made to apply this algorithm in the logarithmic degree regime as well. We skip it here and it may be found for instance in [17]. The algorithm we present and analyze requires the knowledge of the model parameters λ, $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$, although we show in Section 5.4.5, that by a simple modification, we can implement the algorithm even if the model parameters are unknown to the algorithm.

5.4.1 Key Idea behind the Algorithm - Dense Local Interactions

The main and simple idea in our algorithm is that the graph G_n is 'locally-dense' even though it is globally sparse. This is in contrast to sparse Erdős-Rényi based graphs in which the local neighborhood of a typical vertex

'looks like a tree', our graph will have a lot of triangles due to Euclidean geometry. This simple observation that our graph is locally-dense enables us to propose simple pairwise estimators as described in Algorithm 1 which exploits the fact that two nodes 'nearby' in space have a lot of common neighbors (order λ). For concreteness, consider the case when $f_{in}(r) = a\mathbf{1}_{r \leq R}$ and $f_{out}(r) = b\mathbf{1}_{r \leq R}$ for some $R > 0$ and $0 \leq b < a \leq 1$. This means that points at Euclidean distance of R or lesser are connected by an edge in G with probability either a or b depending on whether the two points have the same community label or not. Moreover from elementary calculations, the number of common graph neighbors for any two nodes of G at a distance αR away for some $\alpha < 2$ is a Poisson random variable with mean either $\lambda c(\alpha) R^d (a^2 + b^2)/2$ or $\lambda c(\alpha) R^d ab$ (for some constant $c(\alpha)$ that comes from geometric arguments) depending on whether the two nodes have the same or different community labels. Thus, using a simple strategy consisting of counting the number of common neighbors and thresholding gives a probability of mis-classifying any 'nearby' pair of nodes to be exponentially small in λ. We implement this idea in the subroutine 1 below. Now, to produce the global partition one needs care to aggregate the pairwise estimates into a global partition. Since some pair-wise estimates are bound to be in error, we must identify them and avoid using those erroneous pair-wise estimates (see also Figure 5.1). We achieve this by classifying regions of space B_n as 'good' or 'bad' and then by considering the pair-wise estimates only in the 'good' regions. We prove that if λ is sufficiently large, then the 'good' regions will have sufficiently large volume and hence will

succeed in detecting the communities better than at random.

We summarize our main algorithm below before presenting the formal pseudo-code.

- *Step 1* Partition the region B_n into small constant size cells and based on 'local-geometry' classify each cell as good or bad. This is accomplished in the `Is-A-Good` routine.

- *Step 2* Consider connected components of the Good cells and then in each of them apply the following simple classification rule. We enumerate the nodes in each connected component of Good cells in an arbitrary fashion subject to the fact that subsequent nodes are 'near-by'. Then we sequentially apply the `Pairwise-Classify` Algorithm given in 1.

- *Step 3* Do not classify the nodes in the bad cells and just output an estimate of $+1$ for them.

5.4.2 Notation and Definitions

In this section, we specify the needed notations for describing our algorithm. We will assume that the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfy the hypothesis of Theorem 54. Thus, there exists $0 \leq \tilde{r} < R < \infty$ such that $f_{in}(r) > f_{out}(r)$ for all $r \in [\tilde{r}, R]$. In the rest of this section, we will use the $\tilde{r}$ and R coming from the connection functions.

To describe the algorithm, we need to set some notation. We partition the entire infinite domain $\mathbb{R}^d$ into good and bad regions. However this is just

138

for simplicity and in practice, it suffices to do the partition for the region B_n.
We first tessellate the space $\mathbb{R}^d$ into cubes of side-length $\frac{R}{4d^{1/d}}$ where R is as
above. We identify the tessellation with the index set $\mathbb{Z}^d$, i.e. the cell indexed
z is a cube of side-length $\frac{R}{4d^{1/d}}$ centered at the point $\frac{zR}{4d^{1/d}} \in \mathbb{R}^d$. The subset of
$\mathbb{R}^d$ that corresponds to cell z is denoted by Q_z. Hence the cell indexed 0 is the
cube of side-length $\frac{R}{4d^{1/d}}$ centered at the origin. We now give several definitions
on the terminology used for the $\mathbb{Z}^d$ tessellation and not to be confused with the
terminology for describing the graph G_n. We collect all the different notation
and terminology in this sub-section for easier access and reference.

Definition 61. A set $U \subseteq \mathbb{Z}^d$ is said to be $\mathbb{Z}^d$-**connected** if for every $x, y \in$
U, there exists a $k \in \mathbb{N}$ and $x_1, \cdots x_k \in U$ such that for all $i \in [0, k + 1]$,
$||x_i - x_{i-1}||_\infty = 1$, where $x_0 := x$ and $x_{k+1} := y$.

Definition 62. For any $z \in \mathbb{Z}^d$, denote by $\mathbb{Z}^d$-**neighbors** of z the set of all
$z' \in \mathbb{Z}^d$ such that $||z - z'||_\infty \leq 1$.

Definition 63. For any subset $A \subset \mathbb{Z}^d$ and any $k \in \mathbb{N}$, the k **thickening of**
A is denoted by $\mathbf{L}_k(A) := \cup_{z \in A} \cup_{z' \in \mathbb{Z}^d : ||z - z'||_\infty \leq k} z'$.

Definition 64. For any set $B \subseteq \mathbb{Z}^d$, denote by the set $Q_B := \cup_{z \in B} Q_z$.

Definition 65. Let $\mathcal{Z}(\cdot) : \mathbb{R}^d \to \mathbb{Z}^d$ be the projection function, i.e. $\mathcal{Z}(x) :=$
$\inf\{z \in \mathbb{Z}^d : ||\frac{Rz}{4d^{1/d}} - x||_\infty \leq 0.5\}$. In case, of more than one z achieving the
minimum, we take the lexicographically smallest such z.

Definition 66. For any two points $x, y \in \mathbb{R}^d$, denote by $S_R(x,y) := B(x, R) \cap B(y, R)$, i.e. the intersection of two balls of radius R centered at points x and y.

Definition 67. For any two points $x, y \in \mathbb{R}^d$ such that $||x - y||_2 < R$, define by the two constants $M_{in}(x, y)$ and $M_{out}(x, y)$ as follows.

$$M_{in}(x, y) = \int_{z \in S_R(x,y)} \left(f_{in}(||x - z||) f_{in}(||y - z||) + f_{out}(||x - z||) f_{out}(||y - z||) \right) dz$$

$$M_{out}(x, y) = \int_{z \in S_R(x,y)} \left(f_{in}(||x - z||) f_{out}(||y - z||) + f_{out}(||x - z||) f_{in}(||y - z||) \right) dz.$$

Observe that the definitions of $M_{in}(x, y)$ and $M_{out}(x, y)$ immediately give that

$$M_{in}(x, y) - M_{out}(x, y) =$$
$$\int_{z \in S_R(x,y)} \left(f_{in}(||x - z||) - f_{out}(||x - z||) \right) \left(f_{in}(||y - z||) - f_{out}(||y - z||) \right) dz.$$

Definition 68. For any two points $x, y \in \phi$, denote by $E_G^{(R)}(x, y)$ the number of common graph neighbors of x and y in G which are within a distance R from both x and y.

5.4.3 Algorithm Description in Pseudo Code

We first present two sub-routines in Algorithms 1 and 2 that classify each cell of $\mathbb{R}^d$ to be either Good or Bad. The algorithm is parametrized by $\epsilon \in \left(0, \frac{1}{2}\right)$ which is arbitrary and fixed.

Algorithm 1 Pairwise Classifier

1: **procedure** PAIRWISE-CLASSIFY(i, j, ϕ, G)
2:     **if** $E_G^{(R)}(X_i, X_j) > \frac{\lambda}{2}\left(M_{in}(X_i, X_j) + M_{out}(X_i, X_j)\right)$ **then return** 1
3:     **else**
4: **return** -1
5:     **end if**
6: **end procedure**

In this algorithm, we classify two nodes as in the same partition if the number of common graph neighbors they have exceeds a threshold. The threshold is the average of the expected number of neighbors if the two nodes in consideration are of the same or opposite communities. Such simple tests suffices for our purpose, although one could imagine a more accurate estimator that also takes into account the number of nodes in $S_R(X_i, X_j)$ that do not have any edges to X_i and X_j; or the location labels of the common neighbors.

Algorithm 2 Is A-Good Testing

1: **procedure** IS-A-GOOD(z, G)
2:     **if** $|\phi \cap Q_z| < \lambda(R/4)^d(1/d)(1 - \epsilon)$ **then return** FALSE
3:     **end if**
4:     $\phi^{(z)} := \phi \cap \left(\cup_{z':\|z-z'\|_\infty \leq 1} Q_{z'}\right)$
5:     **for** all $\forall k \geq 1$, and all $X_1, \cdots X_k \in \phi^{(z)}$ **do**
6:         **if** $\prod_{i=1}^k$ PAIRWISE-CLASSIFY$(X_i, X_{i+1}, G) = -1$ **then**   ▷
   Where $X_{k+1} := X_1$
7:             **return** FALSE
8:         **end if**
9:     **end for**
10:     **return** TRUE
11: **end procedure**

To understand the algorithm, we need some definitions which classify

141

Algorithm 3 GBG
___

1: **procedure** MAIN-ROUTINE(G_n, ϕ_n)
2:     Classify each cell in B_n to be either A-Good or A-Bad using subroutine `Is-A-Good`.
3:     Let $\mathcal{D}_1, \cdots \mathcal{D}_k$ be the A-Good $\mathbb{Z}^d$-connected components in B_n.
4:     **for** $l = 1, l \leq k$ **do**
5:         Let $X_{l_1}, \cdots X_{l_{n_j}} \in \phi_n \cap Q_{\mathcal{D}_j}$ be **maximal** and arbitrary s.t $||\mathcal{Z}(X_{l_o}) - \mathcal{Z}(X_{l_{o+1}})||_\infty \leq 1, \forall 1 \leq o \leq n_j - 1$
6:         Set $\hat{\tau}_{l_1}^{(n)} = +1$
7:         **for** $w = 2, w \leq n_j$ **do**
8:             Set $\hat{\tau}_{l_w}^{(n)} = \text{Pairwise-Classify}(l_{w-1}, l_w, \phi_n, G_n)\hat{\tau}_{l_{w-1}}^{(n)}$
9:         **end for**
10:     **end for**
11:     **for** $c = 1, c \leq N_n$ **do**
12:         **if** $\hat{\tau}_c^{(n)} = 0$ **then**
13:             Set $\hat{\tau}_c^{(n)} = +1$
14:         **end if**
15:     **end for**
16:     **return** $\{\hat{\tau}_i^{(n)}\}_{i=1}^{N_n}$
17: **end procedure**
___

cells of $\mathbb{Z}^d$ into Good or Bad depending on the 'local graph geometry'.

Definition 69. A cell Q_z is **A-Good** if

1. $\left| \phi \cap Q_z \right| \geq \max \left(\lambda \left(\frac{R}{4} \right)^d \frac{1}{d} (1 - \epsilon), 1 \right)$; and

2. `Is-A-Good`(z, G) returns TRUE

A cell is called **A-Bad** if it is not A-Good.

The key idea of our simple algorithm lies in the definition of A-Good cells. We classify a cell to be A-Good if there are no 'inconsistencies' in the Pairwise-Estimates. See Figure 5.1 for an example of pair-wise inconsistency due to the `Pairwise-Classify` algorithm. In words, a cell is A-Good, if among the nodes of G that either lie in the cell under consideration or in the neighboring cells, there are no inconsistencies in the output returned by the `Pairwise-Classify` algorithm. Moreover, one can test whether a cell is A-Good or not based on the data (ϕ, G) itself as done in Algorithm 2. Thus, we use the nomenclature of *Algorithm*-Good as A-Good.

The main routine in Algorithm 3 proceeds as follows. In Line 3, we extract out all A-Good connected cells in the spatial region B_n. Suppose that there are k A-Good connected components denoted by $\mathcal{D}_1, \cdots, \mathcal{D}_k$. Our algorithm looks at each connected component independently and produces a labeling of the nodes in them. In Line 5, we enumerate all nodes in any A-Good connected component $\mathcal{D}_l$ as $X_{l_1} \cdots X_{l_{n_l}}$ such that for all $1 \leq o < n_l$, we have $||\mathcal{Z}(X_{l_o}) - \mathcal{Z}(X_{l_{o+1}})||_\infty \leq 1$. Such an enumeration of any A-Good connected
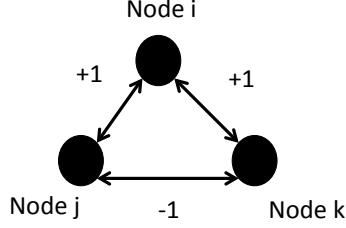
Figure 5.1: An illustration where Pairwise-Classify leads to inconsistency. The values on the edges represent the output of pairwise classify run on the two end points as inputs. In this example it is clear that for at-least one pair $(i, j), (j, k), (k, i)$, the output of pairwise estimate is different from the ground truth.

component is possible since by definition, every A-Good cell is non-empty of nodes. Now, we sequentially estimate the community labels in Line 8 using the `Pairwise-Classify` sub-routine applied on 'nearby' pairs of nodes. In Line 13, we assign an estimate of $+1$, i.e., extract no meaningful clustering for nodes that fall in A-Bad cells. See also Figure 5.2 for an illustration.

5.4.4 Complexity and Implementation

We discuss a simple implementation of our algorithm which takes time of order n^2 to run and storage space of order n. The multiplicative constants here depend on λ. We store the locations ϕ_n as a vector whose length is order λn and the graph G_n as an adjacency list. An adjacency list representation is appropriate since G_n is sparse and the average degree of any node is a constant (that depends on λ). Once we sample the locations ϕ_n, the graph G_n takes time of order n^2 to sample. However, if one represented the locations of nodes

more cleverly, then the sampling complexity could possibly be reduced from n^2. Moreover, since the average degree is a constant, the storage needed is order n. Given the data ϕ_n and the graph G_n, we pre-process this to store another adjacency list where for every vertex, we store the list of all other vertices within a distance of $2R$ from it. This preprocessing takes order n^2 time and order n space. The space complexity is order n since the graph is sparse. Equipped with this, we create a 'grid-list' where for each coordinate of $\mathbb{Z}^d$, we store the list of vertices whose location is in the considered grid cell. This takes just order n time to build. Moreover, since only a constant number of nodes are in any grid cell and the set B_n contains order n cells, the storage space needed for 'grid-list' is order n. Furthermore, since only a constant number of nodes are in a cell, it takes a constant time to test whether a particular cell is A-Good or A-Bad. Thus, to find $\mathbb{Z}^d$ connected components of Good-cells and produce the clustering takes another order nd time where d is the dimension. This gives our algorithm overall a time complexity of order n^2 and a storage complexity of order n.

5.4.5 Practical Implementation if Model Parameters are Unknown

In this section, we provide a simple alternative that can be used to cluster even when the model parameters $f_{in}^{(n)}, f_{out}^{(n)}(\cdot)$ and λ are unknown to the algorithm. Assume for simplicity, that we know an estimate of R such that $\int_{x \in B(0,R)} (f_{in}(||x||) - f_{out}(||x||)) dx > 0$, i.e. the set $\{r \in [0, R] : f_{out}(r) < f_{in}(r)\}$ has non-zero Lebesgue measure. Then, we can change the definition of A-Good

145
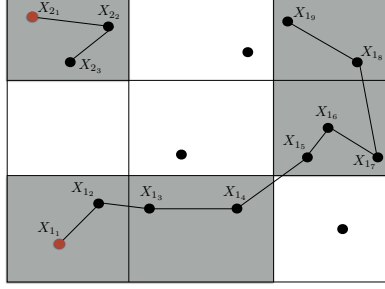
Figure 5.2: An illustration of algorithm 3. In this example, *we do not draw the graph G*, but only show the locations of the nodes. The shaded cells corresponds to A-Good cells and in this example there are two A-Good connected components. In each component, we outline an arbitrary sequence of points $X_{1_1} \cdots X_{1_9}$ and $X_{2_1}, X_{2_2}, X_{2_3}$ that will be used in line 4 of our main Algorithm 3. The lines then represent how we recursively set the community label estimates of the nodes as in line 8 of Algorithm 3. The estimates for the nodes in A-Bad cell is always set to 1.

as follows. For any grid cell $z, z' \in \mathbb{Z}^d$ such that $||z - z'||_\infty = 1$, consider the subgraph $G_{z,z'}$ of G consisting of nodes whose locations lie either in cell z or z'. Consider applying some known partition to the nodes of $G_{z,z'}$, for instance the standard spectral method described in [16]. We can denote the cell z to be A-Good, if the number of points of ϕ in that cell is no smaller than $(1 - \epsilon)$ of the expected value, *and* for every z' in the 1-thickening of z, the partition of the nodes of G_z, when the spectral method is applied to the induced sub-graph $G_{z,z'}$ is identical, i.e. for every z', z'' in the 1 thickening of z, the partition of the nodes of G_z is same whether the spectral method is run on the graph $G_{z,z'}$ or $G_{z,z''}$. Since the spectral method of [16] does not need to know the connection functions, one can use this as an alternative definition of A-Good cell in place of Algorithm 2. The only model information in this alternative

146

implementation required is an estimate of R to perform the tessellation of space. Thus in line 3 of Algorithm 3, we can invoke the test described in this paragraph which does not need knowledge of the connection functions, as opposed to involing Algorithm 2 which does need knowledge of the parameters.

5.5 Lower Bound for Community Detection

The goal of this section is to prove Theorem 50. The central idea is to consider the problem of how well can one estimate whether two uniformly randomly chosen nodes of G_n belong to the same or opposite communities better than at random. This problem is indeed easier than Community Detection which requires one to produce an entire partition of the nodes of G_n. We will show that the natural way to understand the pairwise classification problem is through another problem which we call 'Information Flow through Infinity' which we define in the sequel in Section 5.5.2. Informally, this problem asks whether one can estimate with success probability larger than a half, the community label of any node chosen uniformly at random from G_n, given the graph, the spatial locations **and** the true community labels of all nodes whose spatial locations are *far* away (at infinity) from this chosen node. Subsequently, the core technical argument of this section is to establish an impossibility result for Information Flow from Infinity which we state below in Theorem 75. To aid us in developing the technical arguments, it is instructive to first consider the proof of Proposition 53 (which was stated in Section 5.3), which identifies a special case of connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ when

147

the phase-transition is sharp.

5.5.1 Proof of Proposition 53

Let $R_{in} > R_{out} \geq 0$ be arbitrary and consider the two functions to be $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$. In words, two points of opposite communities are connected if and only if their distance is lesser than R_{out} and two points of the same community are connected if and only if their distance is smaller than R_{in}. In this example, it is clear that for any two points $X_i, X_j \in \phi$, no matter their community labels Z_i and Z_j, if $||X_i - X_j||_2 \leq R_{out}$, then i and j are always connected in G. Similarly, any two points X_i and X_j such that $||X_i - X_j||_2 > R_{in}$ are never connected by an edge in G no matter their community labels Z_i and Z_j. Hence, the *informative* pairs of points in this example are those X_i, X_j such that $||X_i - X_j||_2 \in (R_{out}, R_{in}]$. Moreover, it is immediate that, if $||X_i - X_j||_2 \in (R_{out}, R_{in}]$ and $i \sim_G j$, then $Z_i = Z_j$. On the other hand if $||X_i - X_j||_2 \in (R_{out}, R_{in}]$ and $i \nsim_G j$, then it must be the case that $Z_i \neq Z_j$. For any two points X_i and X_j such that $||X_i - X_j||_2 \in [0, R_{out}] \cup (R_{in}, \infty)$, the presence or absence of an edge is not informative as it is a certain event.

This example motivates the following simple algorithm for Community Detection. Partition the nodes of G_n into $\mathcal{D}_1, \cdots \mathcal{D}_k$ where each component $\mathcal{D}_i$ is a maximal set of nodes $\{X_{i_1}, \cdots X_{i_{l_i}}\}$ of G_n such that for all $j \in [1, l_i]$, we have $||X_{i_{j-1}} - X_{i_j}|| \in (R_{out}, R_{in}]$. In words, we form another graph T_n from the points ϕ_n such that any two nodes i and j of G_n are connected in

148

T_n if and only if $||X_i - X_j|| \in (R_{out}, R_{in}]$. Then $\mathcal{D}_1, \cdots \mathcal{D}_k$ are the connected components of the graph T_n. The algorithm works by considering and labeling each connected component $\mathcal{D}_i$ independently of other components. For each cluster $i \in [1, k]$, estimate the node label of X_{i_1} to be $+1$. Then for every $j \in [2, l_i]$, recursively estimate the node label by the following procedure-

- If $i_{j-1} \sim_{G_n} i_j$ then set $Z_{i_j} = Z_{i_{j-1}}$.

- If $i_{j-1} \not\sim_{G_n} i_j$ then set $Z_{i_j} = -Z_{i_{j-1}}$.

This algorithm considers each of the connected component of T_n enumerated in an arbitrary manner and then labels the nodes in these components. The following very elementary proposition explains when this algorithm will perform well.

Proposition 70. Let $R_{in} > R_{out} \geq 0$ be arbitrary such that $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$. If $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$, then the procedure described above solves Community Detection for this set of parameters.

Note that in view of Theorem 75, Proposition 70 will imply Proposition 53.

Proof. Notice that if $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$, then $f_{in}(r) - f_{out}(r) = \mathbf{1}_{R_{out} \leq r < R_{in}}$. From the properties of the construction of the graph, any two $i \neq j \in \mathbb{N}$ such that $||X_i - X_J|| \in (R_{out}, R_{in}]$ satisfies -

- $Z_i = Z_j$ if $i \sim_G j$

149

- $Z_i \neq Z_j$ if $i \not\sim_G j$.

Hence, it is clear that the algorithm described in the preceding paragraph partitions each cluster $\mathcal{D}_i$, $i \in [1, k]$ exactly in accordance to the ground truth. However, it could be that the estimated signs in each of the connected components $\mathcal{D}_i$ could be flipped from the underlying ground truth and hence the achieved overlap can still be small even though we partition each cluster $\mathcal{D}_i$ accurately. To argue that the overlap achieved by the algorithm is not too small, a sufficient condition is that there exists a unique giant (of size $cn - o(n)$ for some $c > 0$) component of T_n and all other connected components are $o(n)$. Then, we will have by the strong-law of large numbers that the overlap achieved will be c, i.e. the mislabeling in all small components will 'cancel' each other out and in particular cannot drive the overlap of c achieved in the giant component to 0. From the definition of percolation, a unique giant component in T_n exists if and only if $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$ since $T_n \overset{(d)}{=} H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}$.

$\square$

In the sequel, we will generalize the above example to come up with the general lower bound for Community Detection problem.

5.5.2 The Information Flow from Infinity Problem

This problem refers to how well can one estimate the community label of a tagged node of a graph better than at random, given some extra 'information at infinity'. We make this problem precise by posing this question under the

Palm Probability measure $\mathbb{P}^0$. Recall that the Palm measure is the distribution of the graph G obtained by placing an additional node at the origin and equipping it with an independent community label and edges to other existing nodes. For every $r \in \mathbb{R}_+$, denote by $\phi^{(r)}$ and $G^{(r)}$ the point-process and graph, in which every vertex $i \in \mathbb{N}$ (which is at location $X_i \in \mathbb{R}^d$) is equipped with the random variable $Z_i \mathbf{1}_{\|X_i\|_2 \geq r}$. Note that this is not a mark since it is not translation invariant, but is a (measurable) random variable associated with vertex i. In words, we retain the community label marks on nodes of G at a Euclidean distance of r or more from the origin and delete (i.e. set to 0) the community label of those nodes which are located at distances less than r from the origin.

Definition 71. We say *Information Flows from Infinity* if for every $r \in \mathbb{R}_+$ there exists a random variable $\tau_r \in \{-1, +1\}$, measurable (deterministic function) with respect to the observed data $(\phi^{(r)}, G^{(r)})$ and a constant $\gamma > 0$ such that

$$\liminf_{r \to \infty} \mathbb{P}^0[\tau_r = Z_0] \geq \frac{1}{2} + \gamma. \tag{5.4}$$

We say information 'flows' from infinity if we are able to non-trivially estimate the community label at origin, given 'information at infinity'. Note that for each r, there exists algorithms (i.e. τ_r) such that $\mathbb{P}^0[\tau_r = Z_0] > \frac{1}{2}$. However, the non-trivial question is to understand if the limit as $r \to \infty$ is still strictly larger than a half. This definition is similar in spirit to those

considered in Ising models to detect phase-transition for multiplicity of Gibbs states (as in [81] and [288]). We first establish a monotonicity property of this problem and connect it with the Community Detection Problem.

Proposition 72. For every $d \in \mathbb{N}$ and $f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0, 1]$, the limit $\lim_{r \to \infty} \sup_{\tau_r \in \sigma(\bar{G}^{(r)}, \bar{\phi}^{(r)})} \mathbb{P}^0[\tau_r = Z_0]$ exists. Moreover, the map $\lambda \to \lim_{r \to \infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0]$ is non-decreasing.

Note the supremum is over all possible estimators of the community label at origin.

Proof. Denote by $\tilde{\xi}(\lambda, r) := \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0_{\phi}[\tau_r = Z_0]$. Notice that, for each fixed λ and $r' \geq r$, we have $\sigma((\bar{G}^{(r')}, \bar{\phi}^{(r')})) \subseteq \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))$. This follows from the fact that sample path-wise,$(\bar{G}^{(r')}, \bar{\phi}^{(r')})$ is a measurable function of $(\bar{G}^{(r)}, \bar{\phi}^{(r)})$ which is obtained by zeroing all revealed labels in the set $B_r^c \cap B_{r'}$. Hence, the limit in proposition 72 exists.

It remains to prove that $\xi(\lambda) := \lim_{r \to \infty} \tilde{\xi}(\lambda, r)$ is non-decreasing in λ. It suffices to prove that $\tilde{\xi}(\lambda, r)$ is non-decreasing in λ for every r. We show this by using a standard coupling argument used to prove monotonicity of percolation probabilities (for example in Chapter 2, [268]). The basis of the coupling argument is the independent thinning property and Slivnyak's theorem of the PPP and the associated random connection model. These two theorems gives the following two facts. Let (ϕ, G) be a Poisson Point Process of intensity λ and G is the block model graph for some connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ under measure $\mathbb{P}$. Then if each node of G along with its incident

edges are removed independently with probability p, the resulting point process ϕ' is an instance of a PPP with intensity λp and the resulting graph G' is the associated block model graph with the same connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$. Slivnyak's theorem for (ϕ, G) gives that if we place an extra node at origin and equip it with independent community label and edges, the resulting point-process and graph is equal in distribution to (ϕ, G) under the Palm measure $\mathbb{P}^0$.

Thus given a problem instance at intensity λ under measure $\mathbb{P}^0$, we can independently remove nodes of G other than the one at origin with probability p. The resulting graph and the Information Flow from Infinity problem will be that at intensity λp. Thus, the best performance at intensity λ cannot be smaller than that at intensity λp. Since p was arbitrary, we have that the best performance at intensity λ cannot be smaller than that at any intensity $\lambda' \leq \lambda$. In other words, for all $r \geq 0$, $\tilde{\xi}(\lambda', r) \leq \tilde{\xi}(\lambda, r)$. $\qquad \square$

We will need the following classical result on the ergodic property of marks of a stationary point process.

Proposition 73. ([136]) Let $\phi := \{X_1, X_2, \cdots\}$ be a homogeneous PPP with its atoms enumerated in an arbitrary measurable way. Let each atom $i \in \mathbb{N}$ be assigned a translation invariant mark random variable $J_i \in \Xi$ taking values in an arbitrary Borel measurable space $(\Xi, \mathfrak{g})$. Let $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$ be the box of volume n and let $X_{j^{(n)}} \in \phi$ be chosen uniformly at random among the atoms of ϕ that lie in B_n, if any. Then for all $A \in \mathfrak{g}$, the limit $\lim_{n \to \infty} \mathbb{P}[J_{j^{(n)}} \in$

$A]$ exists and satisfies $\lim_{n\to\infty} \mathbb{P}[J_{j^{(n)}} \in A] = \mathbb{P}^0[J_0 \in A]$, where J_0 is the mark of the atom of ϕ at origin under $\mathbb{P}^0$.

The following proposition establishes that Community Detection is harder than Information Flow from Infinity.

Lemma 74. If there exists a Community Detection algorithm (polynomial or exponential time) that achieved an overlap of $\gamma > 0$, then

$$\lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \geq \frac{1+\gamma}{2}.$$

Proof. We will assume that we cannot solve Information Flow from Infinity problem and then conclude that Community Detection is not solvable. More precisely, we will assume that $\lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2}$ and then argue that no Community Detection algorithm can achieve a positive overlap. A Community Detection algorithm achieves an overlap $\gamma > 0$ if when run on the data (G_n, ϕ_n), it produces an output $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ satisfying

$$\frac{|\sum_{i=1}^{N_n} \tau_i^{(n)} Z_i|}{N_n} \geq \gamma, \tag{5.5}$$

with probability $1 - o_n(1)$. Now, an *easier* question corresponds to asking if any two uniformly randomly chosen nodes (with replacement) of G_n belong to the same or opposite community. This question is easier than Community Detection since one way to answer this pairwise question is to first produce a partition of all nodes of G_n and then answer the question for the two randomly chosen nodes. Note that an overlap of γ can be achieved if and only if a fraction $(1 + \gamma)/2$ of the nodes have been correctly classified. Hence the chance that

154

any two uniformly chosen nodes are classified correctly is at-least $(1 + \gamma)/2$. Since we can achieve an overlap of γ with probability $1 - o_n(1)$, the chance that two uniformly randomly chosen nodes of G_n to be correctly classified is at-least $(1 + \gamma)/2 - o_n(1)$. Hence, if we show that the best estimator for answering whether any two randomly chosen nodes from G_n belong to the same or opposite community has a success probability of at-most $\frac{1}{2} + o_n(1)$, then no algorithm exists for solving Community Detection. In the rest of the proof, we will show that if the Information Flow from Infinity cannot be solved, then for every $\epsilon > 0$, the best estimator to estimate whether any two randomly chosen nodes of G_n belong to the same or opposite communities will succeed with probability at-most $\frac{1}{2} + \epsilon + o_n(1)$. This will conclude the proof that no algorithm exists for solving Community Detection in view of the preceding discussion and hence the proof of Lemma 74.

Let $\epsilon > 0$ be arbitrary. Under the assumption that Information Flow from Infinity cannot be solved, there exists a $r > 0$ such that $\sup_{\tau_r \in \sigma((G^{(r)}, \phi^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2} + \frac{\epsilon}{2}$. In words, choose a r such that the Information Flow from Infinity cannot succeed with probability larger than $\frac{1}{2} + \frac{\epsilon}{2}$. Now, let n be large enough such that with probability at-least $1 - \frac{\epsilon}{2}$, we have (i) - G_n has at-least two nodes and (ii) Any two randomly chosen nodes with replacement from G_n denoted as i and j is such that $||X_i - X_j|| > r$. Now, assume that we are on the event that G_n has at-least 2 nodes and two randomly chosen nodes with replacement denoted by i and j are such that $||X_i - X_j|| > r$. From the choice of r, we will argue that conditionally on this

event, the probability that any pairwise estimator correctly tells whether the two nodes i and j are in the same or opposite community will succeed with probability at-most $\frac{1}{2} + \frac{\epsilon}{2}$. This follows since conditionally on X_i and X_j, we can make the pairwise problem *easier* by revealing all community labels of nodes at a distance of larger than r from X_i and asking whether we can now guess the community label at X_i. This will enable us to answer the pairwise question of whether X_i and X_j lie in the same community or not since we will know the true label of X_j when the labels of nodes at distances r or more from X_i are revealed. This now is a problem of finding a mark τ_i of the atom i of ϕ which denotes the best community label estimate of X_i given ϕ, G and all community labels of nodes at a distance of r or more from X_i. Since X_i was an uniformly randomly chosen point from $\phi \cap B_n$, Proposition 73 gives that $\mathbb{P}[\tau_i = Z_i] = \mathbb{P}^0[\tau_0 = Z_0]$. Thus the probability $\tau_i = Z_i$ is bounded from above by $\frac{1}{2} + \frac{\epsilon}{2} + o_n(1)$, as by assumption $\mathbb{P}^0[\tau_0 = Z_0] \leq \frac{1}{2} + \frac{\epsilon}{2} + o_n(1)$. On the complementary event that either $||X_i - X_j|| \leq r$ or if G_n has lesser than 2 nodes, we use the trivial bound that the pairwise estimation is always successful. Hence by the law of total probability, the success probability of the pairwise estimator cannot be larger than $\frac{1}{2} + \epsilon + o_n(1)$. In other words, for every $\epsilon > 0$, there exists a $n_\epsilon < \infty$, such that for all $n \geq n_\epsilon$, the probability that we correctly identify the community membership of two uniformly randomly chosen nodes of G_n is at-most $\frac{1}{2} + \epsilon$.

$\square$

The following is the main technical result on the Information Flow from

Infinity problem.

5.5.3 Main Result on Information Flow from Infinity

Theorem 75. For every $\lambda, f_{in}(\cdot), f_{out}(\cdot)$ and d, the following limit exists and satisfies

$$\lim_{r \to \infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2}\left(1 + \theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d})\right). \quad (5.6)$$

Recall that $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d})$ is the percolation probability of the classical random connection model where any two nodes of ϕ located at $x, y \in \mathbb{R}^d$ are connected by an edge with probability $f_{in}(||x - y||) - f_{out}(||x - y||)$. The supremeum is over all valid estimators of the community label at origin and hence if $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 0$, then there is no estimator that will solve the Information Flow from Infinity problem. In view of Lemma 74, we also get that if $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 0$, then there is no algorithm (polynomial or exponential time) to solve Community Detection. Thus, if we prove Theorem 75, then we will conclude the proof of Theorem 50.

Before presenting the proof of Theorem 75, we illustrate a few example setting where the bound in Equation 5.6 is tight and loose respectively. In view of Lemma 74 and Proposition 53, the following corollary where Equation (5.6) is tight holds.

Corollary 76. For all $\lambda > 0$, $R_1 \geq R_2$, if $f_{in}(r) = \mathbf{1}_{r \leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_2}$

$$\lim_{r \to \infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] = \frac{1}{2}\left(1 + \theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d})\right).$$

In other words, we see that the inequality in Theorem 75 is achieved in certain examples. However, Theorem 75 is not an accurate characterization of the Information Flow from Infinity problem as evidenced in the following example.

Proposition 77. For all $d \geq 2$, if $f_{in}(r) = \min\left(1, \frac{1}{\sqrt{r}} + \frac{1}{r^{d-1/4}}\right)$ and $f_{out}(r) = \min\left(1, \frac{1}{\sqrt{r}}\right)$, the inequality in Equation (5.6) is strict for all values of $\lambda > 0$.

The example in Proposition 77 corresponds to the case when the degree of each node is almost-surely infinite. Thus, $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 1$ in this case. However, using results from [355], one can argue that perfect recovery is impossible in this example, i.e. $\lim_{r \to \infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] < 1$. The key tool, is to see that if perfect recovery were to be possible, then it would be the case that either of the following two pairs of point-process will be mutually singular.

1. The point process formed by the location of those nodes of G that have an edge to the origin and have a community label Z_0 and the point process formed by the location of those nodes of G having an edge to the origin and having a community label of $-Z_0$ are mutually singular.

2. Or, the point process corresponding to the locations of those nodes of G that have a community label Z_0 and do not have an edge to the origin

158

and the point process corresponding to the locations of those nodes of G that have a community label $-Z_0$ and do not have an edge to the origin are mutually singular.

We will argue that in our example, neither is possible by alluding to a theorem from [355], and hence perfect recovery is not possible. We present the complete proof in the Appendix D.

5.5.4   The Information Graph and Proof of Theorem 75

In this section, we generalize the example of the previous section and give a proof of Theorem 75. To do so, we define a general information graph and conclude that if this constructed information graph does not percolate, then one cannot solve the Information Flow from Infinity problem.

We denote by I the information graph whose vertex set is ϕ. The random graph I is constructed just based on the positions of the points and the random elements $\{\{U_{ij}\}_{j>i}\}_{i\in\mathbb{N}}$. Recall that the graph G was built by connecting any two points $i < j \in \mathbb{N}$ if $U_{ij} \le \mathbf{1}_{Z_i=Z_j} f_{in}(||X_i - X_j||) + \mathbf{1}_{Z_i\neq Z_j} f_{out}(||X_i - X_j||)$. Using the same random elements, we connect any $i < j \in \mathbb{N}$ by an edge in graph I if $U_{ij} \in [f_{out}(||X_i - X_j||), f_{in}(||X_i - X_j||)]$. We denote by $i \sim_I j$ the event that points i and j are connected by an edge in I. Hence the graphs I and G are coupled and built on the same probability space using the same set of random elements. For each $i \in \mathbb{N}$, we denote by $V_I(i) \subseteq \mathbb{N}$ the random subset of the nodes contained in the connected component of node i in graph I. Note that the information graph $I \stackrel{(d)}{=} H_{\lambda, f_{in}(\cdot)-f_{out}(\cdot),d}$, i.e. the I graph we

159

constructed is equal in distribution to the graph of the Poisson Random Connection model with vertex set forming a PPP of intensity λ and connecting any two vertices at distance r away with probability $f_{in}(r) - f_{out}(r)$ independently of everything else. This equality in distribution follows from the fact that $\{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N}}$ is an i.i.d. uniform $[0,1]$ sequence. The following structural lemma justifies the term information graph.

Lemma 78. From the way we have coupled the construction of G and I, we have

- If $i \sim_I j$ and $i \sim_G j$, then $Z_i = Z_j$.

- If $i \sim_I j$ and $i \nsim_G j$, then $Z_i \neq Z_j$.

Proof. This follows from the following construction of G and I as follows.

- $i \sim_G j$ if and only if $U_{ij} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i = Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i \neq Z_j}$.

- $i \sim_I j$ if and only if $U_{ij} \in [f_{out}(||X_i - X_j||), f_{in}(||X_i - X_j||)]$.

- $\forall r \geq 0, \ f_{in}(r) \geq f_{out}(r)$.

The lemma follows since the $\{U_{ij}\}_{0 \leq i < j}$ are the same with which we build both the random graphs G and I. $\qquad\square$

We can iterate the above lemma from edges to connected components of I which forms a crucial structural lemma.

Lemma 79. For all $j \in \mathbb{N}$, conditional on G, ϕ, I, there are exactly two possible sequences $(Z_k)_{k \in V_I(j)}$ which are complements of each other that are consistent in the sense of Lemma 78 with the observed data G, ϕ and I.

The proof of this follows from Lemma 78 and an induction argument. The proof can be found in Appendix D.4. We now present the main probabilistic observation in the sequel in Lemma 80 which essentially states that the community labels on disconnected components are independent.

Lemma 80. For all $\lambda > 0$, on the event $\{|V_I(0)| < \infty\}$,

$$\mathbb{P}^0 \left[Z_0 = +1 \middle| G, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, \phi, \{Z_i\}_{i \in V(I_{(0)})^c} \right] = \frac{1}{2} \text{ a.s.}$$

Proof. From Lemma 79, we know that conditionally on ϕ, G, I, there are exactly two possible sequences $\{Z_k\}_{k \in V_I(0)}$ that are consistent with the observed data in the sense of Lemma 78. Denote these two sequences by $\mathbf{s}$ and $\mathbf{s}^c$. It suffices to show that conditionally on $\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, G$ and $\{Z_k\}_{k \in V_I^c(0)}$, the two sequences $\mathbf{s}$ and $\mathbf{s}^c$ are equally likely. We will denote by g the realization of the random graph G. To conclude the lemma, we use Bayes' conditional rule as follows.

$$\mathbb{P}^0_\phi[(Z_k)_{k \in V_I(0)} = \mathbf{s} | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}, G = g]$$

$$= \frac{\mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}, (Z_k)_{k \in V_I(j)} = \mathbf{s}]}{\mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}]} \tag{5.7}$$

$$\mathbb{P}^0_\phi[(Z_k)_{k \in V_I(0)} = \mathbf{s} | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}]$$

$$\overset{(a)}{=} \frac{1}{\sum_g \mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}]} \left(\frac{1}{2}\right)^{|V_I(0)|} \quad \text{a.s. on } \{|V_I(0)| < \infty\} \tag{5.8}$$

$$\overset{(b)}{=} \frac{1}{2} \quad \text{a.s. on the event } \{|V_I(0)| < \infty\} \tag{5.9}$$

The first equality follows from rewriting the events using Baye's conditional rule. In the rest of the proof, we justify steps (a) and (b). We prove the equalities and also justify that one can apply conditional Baye's rule without worrying about the 0 by 0 situation almost-surely.

Note that conditionally on $\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, \{Z_k\}_{k \in \mathbb{N}}$, the graph G is fixed and deterministic. Thus, the numerator in step (a) is 1 almost-surely. This follows from Lemma 79 which states that g is consistent with the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)})$ if $(Z_k)_{k \in V_I(0)} = \mathbf{s}$ or $\mathbf{s}^c$. Furthermore, the process $(Z_k)_{k \in \mathbb{N}}$ is an i.i.d. sequence independent of everything else. Hence, given any random finite subset $A \in \mathbb{N}$ independent of $(Z_k)_{k \in \mathbb{N}}$, the labels $(Z_k)_{k \in A}$ are uniform over $\{-1, 1\}^{|A|}$. Now, since $|V_I(0)| < \infty$, and $V_I(0)$ is a function of $(\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}})$ which is independent of $(Z_k)_{k \in \mathbb{N}}$, it follows that, on the event $\{|V_I(0)| < \infty\}$,

$$\mathbb{P}^0_\phi[(Z_k)_{k \in V_I(0)} = \mathbf{s} | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}] = \left(\frac{1}{2}\right)^{|V_I(0)|} \quad \text{a.s.} \tag{5.10}$$

Moreover, the above expression is non-zero almost surely since $|V_I(0)| < \infty$. This justifies step (a). To conclude the proof, it suffices to show that on the event $\{|V_I(0)| < \infty\}$,

$$\sum_g \mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)}] = 2 \left(\frac{1}{2}\right)^{|V_I(0)|} \quad \text{a.s .} \quad (5.11)$$

This will conclude the proof by noticing that the above expression is non-zero almost-surely.

Observe that the summation in Equation (5.11) is over the various community labels $(Z_k)_{k \in V_I(0)}$. Thus, the summation is over the $2^{|V_I(0)|}$ different choices for $(Z_k)_{k \in V_I(0)}$. However, given ϕ and $\{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}$, one can construct the I graph. Then, Lemma 79 states that the total number of possible choices for the labels $(Z_k)_{k \in V_I(0)}$ is now only two, which we denoted by $\mathbf{s}$ and $\mathbf{s}^c$ in this proof. However, again from Lemma 79, conditionally on those two sequences, the graph constructed from the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)},$ $(Z_k)_{k \in V_I(0)} = \mathbf{s})$ and from the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)},$ $(Z_k)_{k \in V_I(0)} = \mathbf{s}^c)$ is g, the observed graph. Hence, the proof of the claim follows from Equation (5.10).

$\square$

The following is an immediate corollary of the definition of conditional expectation.

Corollary 81. For all events $A \in \sigma(G, \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V^c_I(0)})$, we have $\mathbb{E}^0[\mathbf{1}_E \mathbf{1}_A \mathbf{1}_{Z_0=+1}] = \mathbb{E}^0[\mathbf{1}_E \mathbf{1}_A \mathbf{1}_{Z_0=-1}] = \frac{1}{2}\mathbb{E}^0[\mathbf{1}_E \mathbf{1}_A]$, where E is the event that $V_I(0)$ is finite.

5.5.5 Proof of Theorem 75

We are now ready to conclude the proof of Theorem 75. Notice that since $\tau_r \in \{-1,+1\}$, we can represent it as $\tau_r = \mathbf{1}_A - \mathbf{1}_{A^c}$, for some $A \in \sigma((G^{(r)}, \phi^{(r)}))$. Hence, we have

$$\sup_{\tau_r \in \sigma((G^{(r)},\phi^{(r)}))} \mathbb{P}^0_\phi[\tau_n^{(\delta)} = Z_0] = \sup_{A \in \sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0_\phi[\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1}]. \quad (5.12)$$

For every $m \in \mathbb{N}$, denote by E_m the event that $C_I(0) \subseteq B_m$, i.e. the event that the connected component of the point at the origin in I is contained in the set B_m. The sets E_m are non-decreasing. Moreover, from the definition of percolation, $\mathbb{P}^0[\cup_{m\in\mathbb{N}} E_m] = \lim_{m\to\infty} \mathbb{P}^0[E_m] = 1 - \theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d})$. Let $r \in \mathbb{R}$ be arbitrary, and condition on the event E_r. We have,

$$\sup_{A \in \sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1}] =$$

$$\sup_{A \in \sigma(((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_{E_r}(\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1})] + \mathbb{E}^0[\mathbf{1}_{E_r^c}(\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1})]$$

$$\leq \sup_{A \in \sigma((((G^{(r)},\phi^{(r)})))} \mathbb{E}^0[\mathbf{1}_{E_r}(\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1})] + \mathbb{P}^0[E_r^c]$$

$$\overset{(a)}{\leq} \sup_{A \in \sigma((G^{(r)},\phi^{(r)},\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}})} \mathbb{E}^0[\mathbf{1}_{E_r}(\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1})] + \mathbb{P}^0[E_r^c]$$

$$\overset{(b)}{\leq} \sup_{A \in \sigma(G,\phi,\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}},(Z_k)_{k\in V_I^c(0)})} \mathbb{E}^0[\mathbf{1}_{E_r}(\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1})] + \mathbb{P}^0[E_r^c]$$

$$\overset{(c)}{=} \sup_{A \in \sigma(G,\phi,\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}},(Z_k)_{k\in V_I^c(0)})} \frac{1}{2}\mathbb{E}^0[\mathbf{1}_A \mathbf{1}_{E_r}] + \frac{1}{2}\mathbb{E}^0[\mathbf{1}_{A^c} \mathbf{1}_{E_r}] + \mathbb{P}^0[E_r^c]$$

$$= \frac{1}{2}\mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^c].$$

Step (a) follows from enlarging the sigma algebra over which we are searching for a solution. Step (b) follows from the fact that on the event E_r, $V_I(0) \subseteq B(0,r)$. Thus, revealing more labels will only preserve the inequality. Step (c) follows from Corollary 81. Now, since the sets E_r are non-decreasing , we get the theorem by taking a limit as r goes to infinity on both sides, i.e.

$$
\begin{aligned}
\lim_{r \to \infty} \sup_{A \in \sigma((G^{(r)}, \phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_A \mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^c} \mathbf{1}_{Z_0=-1}] &\leq \lim_{r \to \infty} \frac{1}{2} \mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^c] \\
&= \frac{1}{2} \mathbb{P}^0[E] + \mathbb{P}^0[E^c] \\
&= \frac{1}{2} (1 + \theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d})).
\end{aligned}
$$
(5.13)

The limit on the LHS exists from Proposition 48 and the limit on the RHS exists since E_r are non-decreasing events.

5.6 Conclusions and Open Problems

In this chapter, we introduced the problem of community detection in a spatial random graph where there are two equal sized communities. We studied this problem in both the sparse and non-sparse regime. Our main technical contributions in the sparse graph case are in identifying the problem of Information Flow from Infinity and connecting that with the Community Detection problem and giving a simple lower bound criterion. For developing the algorithm, we noticed that a spatial graph is sparse due to the fact that all interactions are dense, but localized which is starkly different from the reason why an Erdős-Rényi graph is sparse. We leveraged this difference to

propose an algorithm for community detection by borrowing further ideas from dependent site percolation processes. In the Exact-Recovery setting, we give a lower bound that we conjecture to be tight. However, this is just a first step and there are a plenty of open questions just concerning the model we introduced.

1) Is Weak-Recovery and Information Flow from Infinity equivalent ? - Here, we proved that weak-recovery was harder than Information Flow from Infinity. However, our algorithm and its analysis showed that it can solve Community Detection whenever it can solve Information Flow from Infinity. Thus a natural question is whether these two problems undergo a phase-transition at the same point ? Moreover is there a relation between the optimal overlap achievable in Community Detection and the optimal success probability of estimating the community label of the origin in the Information Flow from Infinity problem ?

2) Is the optimal overlap in weak-recovery monotone non-decreasing in λ *?* - We see in the proof of Proposition 48 that the optimal success probability of correctly labeling the origin in the Information Flow from Infinity problem is monotone in λ. However, for Community Detection, we only established that solvability is monotone and not the optimal overlap achievable.

3) Can one resolve Conjecture 59 to identify the critical phase-transition point for Exact-Recovery - This conjecture is reminiscent of the local to global phenomenon consistently observed in various random graph models ([16], [302],[286]). In these settings, an obvious local condition, i.e., there being no flip-bad

166

node, also turns out to be sufficient for exact-recovery. Establishing such a result in our model will help us obtain a better understanding of our random graph model and also may aid in improved algorithms for practical situations.

3) More than 2 Communities - In this thesis, we focused exclusively on the case of two communities in the network, and an immediate question is that of 3 or more communities. In the symmetric case where the connection function is $f_{in}^{(n)}(\cdot)$ within communities and $f_{out}^{(n)}(\cdot)$ across communities, a simple adaptation of our algorithm can give a sufficient condition in both the sparse and logarithmic degree regime, although will be sub-optimal. Our lower bound technique in the sparse regime can also be applied in the setting of many communities (see Theorem 2 in [19] for example) thereby establishing the *existence* of a non-trivial phase transition for any number of communities in the symmetric setting. But the open question is to identify examples similar to Proposition 53 where the phase transition for weak-recovery can be tight. A quest for such examples can possibly lead to better understanding of even the 2 community case considered in this thesis. Moreover, unified algorithmic techniques capable of handling non-symmetric case also is of interest since our algorithm does not generalize in a straight forward way to the non-symmetric setting. As far as the logarithmic degree regime, our lower bound framework can easily extend using the same large-deviations result for Poisson hypothesis testing developed in [20]. However, an algorithm achieving this threshold in the logarithmic regime with multiple communities is not yet known.

4) Characterization of the Phase-Transition for Weak-Recovery - An

167

obvious but harder question is whether one can characterize if not compute the exact phase-transition for either Community Detection or Information Flow from Infinity. We show that our lower bound is capturing the phase-transition only in very specific cases and may not be tight in general due to corner cases similar to Proposition 77. We also have no reason to believe that our algorithm is optimal in any sense. Thus, a structural characterization of the phase-transition is still far from being understood.

5) Computational Phase-Transition - Another aspect concerns the possible gaps between information versus computation thresholds. Is there a regime where Community Detection is solvable, but no polynomial (in n) time and space algorithms that operate on (G_n, ϕ_n) are known to exist ?

6) Estimating the Model Parameters - How to efficiently estimate $f_{in}(\cdot)$ and $f_{out}(\cdot)$ from the data of just the graph G and the spatial locations ϕ.

Chapter 6

Haplotype Assembly and Community Detection

In this chapter we discuss how to use the insights gained from the mathematical study of Euclidean Community Detection in the previous chapter, to aid in design of practical algorithms for Haplotype Phasing. This chapter is organized as follows. In Section 6.1, we introduce the problem of Haplotype Phasing and its importance in computational biology. In Section 6.2 we precisely state the mathematical problem formulation of haplotype assembly along with the key performance metrics. In section 6.3, we describe our algorithm `ComhapDet`, a community detection based algorithm to solve the problem. We evaluate the performance of our algorithm empirically and compare it against other state of the art methods in Section 6.4.

6.1   Introduction and Background

Haplotype Phasing problem consists of reconstructing individual 'haplotypes' or DNA sequences from a collection of noisy measurements or 'reads'. Every species has an associated DNA sequence, which for the purposes of this note can be considered to be a sequence of $\{A, C, G, T\}$ alphabets called 'bases'

169

or 'alleles'. The DNA sequence however is known to exhibit small variations across different individuals of an organism, even though the DNA sequence is 'approximately' identical across all members of an organism. Amongst the variations of the DNA sequence between individuals, an important class is what is known as the Single Nucleotide Polymorphisms (SNPs). The ordered list of alleles or the bases at these SNP variant positions is known as *haplotypes*. In humans for instance, each individual has two copies of the DNA string, that are identical except at SNP, where the bases on the two strings are different. Moreover, different individuals have different pair of bases at a SNP position. Humans are thus referred to as *diploid*, since there are two roughly identical strings of DNA with variations at SNP positions. Many plants on the other hand are *polyploid* species, i.e., each individual has multiple DNA strings associated with them, which are all almost identical except at SNP positions, where the bases in the different strings are different. For instance the most commonly grown Potato crop (*Solanum Tubersoum*) is a tetraploid organism, i.e., each organism has 4 roughly identical DNA strings, with variations at SNP positions. Technicaly speaking however, each organism has multiple such collection of DNA strings. For example, humans have 22 pairs of DNA sequences, one for each 'chromosome'. Technically, this implies that a human individual, has 22 different instances of the haplotype phasing problem, one for each chromosome.

We briefly describe the haplotype assembly problem here before giving a precise mathematical description of the problem in Section 6.2. The task of

170

haplotype assembly consists of identifying K different strings (the haplotype sequences) denoted as $S_1, \cdots, S_K$, with each string position taking an alphabet in {A,C,G,T}. The K strings are such that at no position all the K strings have an identical alphabet. In reality, these K strings are the 'ground-truth' alphabets at the SNP positions in a DNA sequence. The K strings are to be deduced from measurements of short *sequencing reads*, where each read originates from one of the K haplotype sequences, and reveals the bases in a subset of the positions. We assume that the positions of each reads are known, i.e., the alphabet indices. However, for each read, it is not known from which of the K haplotype sequence it originates from. The task of haplotype assembly then consists in reconstructing the K strings from a collection of such short reads.

In practice, the K strings consists only of the DNA sequence sampled at the SNP positions. Since the SNP positions are estimated separately (and hence we assume to be known exactly in this chapter), one can only restrict attention to those positions on the DNA sequence. Thus, from the discussions in the previous section, $K = 2$ for humans and $K = 4$ for potato. The assumption of known location positions is also very reasonable in practice. This is so, as it is now reasonable to assume that DNA sequences of organisms of interests has been 'sequenced', i.e., can be assumed to be known. Thus, given a read, one can map it to this genome to infer from which locations the read originates from on the reference genome. However, as the SNP locations are also known, one can after mapping the read to a reference, identify the

SNP positions that a particular read covers.

In the absence of measurement noise, this is a trivial task, as we know that at every position, not all of the K strings have an identical base. Thus, one can partition the set of reads unambiguously as originating from one of th K different haplotypes, and this can be used for haplotype assembly. Unfortunately however, sequencing is erroraneous, and the state of the art sequencing platforms have an error in the magnitude of $10^{-3} - 10^{-2}$. Thus, in the presence of errors, it is no longer obvious as to how to assign whether a read originates from a particular haplotype, and one needs to develop a computational framework for haplotype assembly, which is the central subject of the present chapter.

6.2 Problem Formulation

Let m be the length of the haplotype sequences and n to denote the total number of paired-end read measurements and k denote the ploidy. Thus, there are k different strings, or haplotype sequences to be estimated, each of length m, with the cardinality of the alphabet set being a. The alphabet cardinality a is also refered to the *allelic* of the problem, i.e., if $a = 2$, it is refered to as the bi-allelic case. In most practical applications either $a = 4$, corresponding to the 4 bases of A, C, G, T, or $a = 2$, as in the human haplotype problem case. These strings (or haplotype sequences) are denoted by $s_i[l]$, where for each $i \in \{1, \cdots, m\}$ and $l \in \{1, \cdots, k\}$, where in the tetra-alleleic case, (i.e, the number of alleles is 4), we have $s_l[i] \in \{A, C, G, T\}$. In the rest of the chapter,

172

we denote by the haplotype positions $\{1, \cdots, m\}$ also by the term *sites*. Each measurement, denoted as r_u, for $u \in \{1, \cdots, n\}$ consists of choosing a string $v_u \in \{1, \cdots, k\}$, a set of sites $\mathfrak{I}_u \subset \{1, \cdots, m\}$, and $\{\tilde{s}^{(u)}[i]\}_{i \in \mathfrak{I}_u}$, where for each $i \in \mathfrak{I}_u$, $\tilde{s}^{(u)}[i] \in \{A, C, G, T\}$ is a 'noisy copy' of the ground truth $s_{v_u}[i]$. For each measurement $u \in \{1, \cdots, n\}$, we observe the set of positions $\mathfrak{I}_u$ and the noisy string values $\{\tilde{s}^{(u)}[i]\}_{i \in \mathfrak{I}_l}$, but not the string index v_u, from which the measurement was made. The goal of Haplotype phasing then is, to recover the k ground-truth strings, from which the measurements were observed.

We also assume that the set of m haplotypes and n reads, form a single *contiguous block*. Formally, given the set of m haplotypes and n reads, we say it forms a *contiguous block* if for any two positions $i, j \in \{1, \cdots, m\}$, we have a set of reads $r_1, \cdots, r_k$, such that for every $i \in \{1, \cdots, k-1\}$, the reads r_i and r_{i+1} overlap in at-least one site, and r_1 overlaps with site i and r_k overlaps with site j. Notice that since the problem definition is agnostic to how we label the strings from $\{1, \cdots, k\}$, any haplotype phasing algorithm can only hope to recover the k strings, upto a permutation of the labels. Thus in particular, for two disjoint blocks of haplotype, where there is no read bridging the two blocks, then one can only hope to recover the k strings in the two blocks individually, but there is no way to ascertain which of the k estimated strings in one block is from the same source as one of the k estimated strings in another block. Thus, without loss of generality, we assume that the set of haplotypes m form a contiguous block of reads. Indeed, if this were not the case, then we can pre-process the reads and split the problem into many instances, where

each instance consists of a single contiguous haplotype block that needs to be estimated.

In this chapter, we are interested in the case when the reads/measurements form what are known as *paired-ended reads.* Formally, this implies that each measurement $l \in \{1, \cdots, n\}$, is such that the set of sites covered by read l has *two contiguous* blocks. More precisely, we assume that every read $l \in \{1, \cdots, n\}$ is such that there exists $i_1^{(l)}, j_1^{(l)}, i_2^{(l)}, j_2^{(l)} \in \{1, \cdots, m\}$, such that $\mathcal{I}_l = \{i_1^{(l)}, i_1^{(l)} + 1, \cdots, i_1^{(l)} + j_1^{(l)}\} \cup \{i_2^{(l)}, i_2^{(l)} + 1, \cdots, i_2^{(l)} + j_2^{(l)}\}$.

Recovery Goals and Performance Metrics

The main performance metrics used to benchmark haplotype assembly algorithms is what are known as the Correct Phasing Rate (CPR) and the Minimum Error Correction (MEC) score (see eg. ([117],[194], [177])). The CPR measures the discrepancy between the reconstructed haplotype $\hat{s}_1, \cdots, \hat{s}_k$ and the original ground truth $s_1, \cdots, s_k$, and is defined as,

$$\text{CPR} = \frac{1}{m} \sum_{i=1}^{m} \max_{\pi \in \mathcal{S}_k} \prod_{j=1}^{k} \mathbf{1}_{s_j[i]=s_{\pi(j)}[i]}, \tag{6.1}$$

where $\mathcal{S}_k$ is the set of all permutations of $\{1, \cdots, k\}$. Observe that this is a much more stringent notion of recovery as compared to that used in [193] and [289], where the Reconstruction Rate, which we denote as M-CPR to abbreviate Modified CPR. This metric, is defined by

$$\text{M-CPR} = \max_{\pi \in \mathcal{S}_k} \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} \mathbf{1}_{s_j[i]=s_{\pi(j)}[i]}. \tag{6.2}$$

Observe that in the case of diploid with binary alphabets, both CPR and M-CPR are identical. However, in the polyploid case, where the alphabet size is 3 or more, the inequality CPR $\leq$ M-CPR holds, since for all sets $X_1, \cdots, X_k$, $\prod_{j=1}^{k} \mathbf{1}_{X_j} \leq \sum_{j=1}^{k} \mathbf{1}_{X_j}$.

However, we cannot design an algorithm to directly minimize this objective, as the ground truth is unknown. A commonly used metric in practice is the MEC which can be computed directly from the observed data and the reconstructed output. The MEC score of an algorithm is defined as follows.

$$\text{MEC} = \sum_{j=1}^{n} \min_{p \in \{1, \cdots, k\}} \sum_{i=1}^{m} \mathbf{1}_{\text{Read } j \text{ covers position } i} \mathbf{1}_{\tilde{s}^{(j)}[i] \neq s_p[j]}.$$

In applications where there is no underlying ground truth known, the MEC score serves as a proxy to measure the performance of haplotype assemble methods. Indeed, popular haplotype assembly algorithms, such as [193], [63], try and directly minimize the MEC score, by relaxations of the underlying combinatorial optimization problem. Contrary to this popular approach, in this paper, we do not directly try to minimize the MEC as the problem of minimizing MEC is known to be a NP-Hard combinatorial problem ([226],[87]). More importantly, the MEC score only serves as a proxy to the required performance metric. Instead, we leverage the *typical structure* in the data, and the fact that the noise in measurements are introduced through randomness, rather than the measurements being corrupted by an adversarial source, to design a randomized algorithm by posing haplotype phasing as a spatial graph clustering problem.

The key parameters that impact the performance are *coverage*, *error rate* and *read length*. Formally, the read-length r is defined to be the average of all $j_1^{(l)}$ and $j_2^{(l)}$, i.e., the read-length

$$r = \frac{1}{2n} \sum_{l=1}^{n} (j_1^{(l)} + j_2^{(l)}).$$

We define the coverage $\mathcal{C}$ as the average number of reads that cover a single base in a haplotype, i.e., $\mathcal{C} = \frac{2nr}{km}$. The interpretation is that there are a total of n reads, each having on average $2r$ bases. Thus, the total average number of bases recored is $2nr$. These bases *cover* a total of km bases, i.e., we have k haplotypes, each of m bases long. The error-rate p is the average error rate in the measurement process, i.e., the fraction of bases recored by the reads that are an incorrect copy of the underlying ground truth. Notice that this quantity cannot be defined only based on the observed samples, but is known by other methods, for instance one knows accurate estimates and bounds on the error rates of many of the high throughput sequencers. It is standard practice in the study of haplotype assembly to characterize and benchmark performance of algorithms using either the achieved MEC when the ground truth is unknown and the CPR in simulated data when ground truth is known. We shall characterize the performance of our algorithm, when the problem parameters, namely ploidy and the alphabet size, and the measurement characteristics, namely the coverage, read-lengths and error-rates vary.

Main Contributions

This chapter introduces a *spatial point process representation* of the paired-end reads, which is subsequently used to reconstruct the underlying strings. More precisely, we advocate that a good description of the paired-end read data is to represent each read as having a spatial coordinate corresponding to the starting indices of the two read blocks, in addition to the actual bases recorded by a read. Equipped with such a representation of the data, we then construct a weighted graph among the set of reads, by placing an appropriately weighted edge between two reads . This weighting mechanism ensures that if two reads belong to the same haplotype, then they will likely receive a large positive weight, while if the two reads belong to different haplotypes, then then will likely receive a large negative weight. We then cast the Haplotype Assembly problem as an Euclidean Community Detection problem in the sense of [337], where the community labels of a node (or a read) indicates the haplotype from which it comes from. We find in our experiments, that such a 'spatial' embedding of the problem greatly improves both the accuracy of polyploid phasing and the run time complexity.

6.3   The Haplotype Assembly Algorithm

The algorithm we propose is based on identifying a simple connection between the aforementioned haplotype reconstruction problem and euclidean community detection. Although, such a connection was observed for the special case of single-ended read and the diploid haplotype phasing problem in

[116], the extension to paired-end reads and in particular to the case of phasing polyploids was not known. We provide a unified framework to apply algorithms developed for Euclidean Community detection (for ex. [337]) to both diploid and polyploid haplotype phasing problems.

Connection with Community Detection

In order to invoke a connection with spatial community detection, we pre-process the n paired-end measurements into a graph G on n nodes, with each node representing a measurement. For any two reads $u, v \in [n]$ with $u \neq v$, denote by the intersection of sites at which the two measurements occur as $\mathcal{I}_u \cap \mathcal{I}_v := \{l_1, \cdots, l_q\}$ with $q = 0$ implying this set is empty. More precisely, each l_i, for $i \in \{1, \cdots, q\}$ is a position from 1 through m of the haplotype sites, which both the reads u and v cover. If $q = 0$, it implies that reads u and v cover disjoint set of sites, and in this case, there is no edge between u and v in the graph G. If the two reads u and v, have at-least one site in common, then we place an edge between reads u and v with weight $w_{uv} := \frac{1}{q} \sum_{h=1}^{q} \left(\mathbf{1}_{\tilde{s}^{(u)}[l_h] = \tilde{s}^{(v)}[l_h]} - \mathbf{1}_{\tilde{s}^{(u)}[l_h] \neq \tilde{s}^{(v)}[l_h]} \right)$. In words, the weight between reads u and v, look at what fraction of the sites in which both u and v take a measurement agree minus the fraction of common sites for u and v in which they disagree. *Typically* that the weights $w_{uv} \in [-1, 1]$ for all $u, v \in [n]$. Such a weighting scheme ensures that if w_{uv} is positive and large, then it is likely that measurements u and v are generated from the same string, while if w_{uv} is negative and large in magnitude, then it is likely that measurements u and v are

178

generated from different strings. The intuition for the facts that large positive weights indicate reads come from the same haplotype while large negative weights indicate reads come from different haplotypes can be understood by examining the *typical structure* of the polyploid phasing problem. Observe that if the SNP positions were called accurately, i.e., all of the m haplotypes to be phased were 'true', then it would be the case that in any location $i \in \{1, \cdots, m\}$, not all strings $s_1, \cdots, s_k$ will have identical bases, i.e., the set of locations $\{i \in \{1, \cdots, m\} : s_1[i] = \cdots = s_k[i]\} = \emptyset$. Since sequencing errors are 'typically' small, it is thus the case that if two reads covering the same site have different values, then it is *likely* that they come from different haplotypes.

Additionally, we equip each node $u \in [N_n]$ of this graph (i.e., each paired-end read) two labels - a *community label* $Z_u \in [k]$ and a *spatial label* $X_u \in [m]^2$. The community label of a measurement is the string index or the haplotype index from which that read originated, and the spatial label of a measurement $u \in [n]$ is $(i_1^{(u)}, i_2^{(u)})$, i.e., the indices from which the two contiguous substrings of the read start. If the reads were to be perfect paired-end measurements, then every read would have exactly two contiguous set of sites which it covers, and their starting points form a reads spatial label. However, many times, reads are imperfect, i.e., a single read may have either a single or more than 2 contiguous fragments. In the case that reads have a single fragment, then the spatial location of that reads is to be taken on the diagonal in $[n]^2$, i.e., the starting site index for the supposedly two contiguous fragments are the same. In case of a read having more than 2 fragments, there

179

are several alternatives. We may either split the read into multiple reads, such that each of them have at-most two contiguous fragments. Else, we may pick two of the many starting contiguous points from the read in a suitable fashion as a starting spatial label.

Note that, we get to observe the graph G, after pre-processing the observed data, and the spatial labels $(X_u)_{u \in [n]}$ but not the community labels $(Z_u)_{u \in [n]}$. We defer a discussion of the computational complexity involved in this pre-processing to construct the graph G the sequel, where we show that one can exploit the structure in the data to reduce complexity of this pre-processing step from the naive order n^2 to roughly linear in n. Equipped with such a representation of the data, we reduce the task of haplotype assembly into first performing community detection on the graph G, where the nodes have an additional location label, to cluster the reads as originating from the k different strings. Subsequently, for all $i \in [m]$ and $l \in [k]$, we estimate $\hat{s}_l[i]$ to be the majority alphabet indicated by all reads covering site i and are estimated to originate from string l by the community detection algorithm. Before presenting the entire pseudo-code, we find it instructive to summarize our algorithmic pipeline here.

1. We tessellate the grid $[n]^2$ into smaller *overlapping* boxes, denoted by $(B_{x,y})_{1 \leq x \leq \tilde{n}, 1 \leq y \leq \tilde{n}}$. Here $\tilde{n} < n$ is a parameter which we will set and each $B_{x,y} \subset [n]^2$. This tessellation is such that, each grid point $u \in [n]^2$ will belong to multiple boxes $B_{x,y}$, as the boxes are overlapping.

2. For each box $B_{x,y}$, denote by $H_{x,y}$ to be the subgraph of G containing nodes whose spatial locations lie in $B_{x,y}$. The nodes of $H_{x,y}$ are all clustered independently into k communities.

3. The community estimates in different boxes are *synchronized* to obtain a global clustering estimate from such spatially-local clustering estimates. Since each grid point is present in multiple boxes, a read gets many estimates for its community, each of which adds an 'evidence' to the label of the node. This scheme has a natural 'error-correcting' mechanism, since it is less likely for a node to be misclassified in the majority of the boxes it lies in, as opposed to any one particular box.

4. Finally, for all $j \in [k]$ and $i \in [m]$, we estimate $\hat{s}_j[i]$ to be the majority among the 4 alphabets as indicated by those reads that cover site i and are estimated to belong to string j in the above clustering step.

Euclidean Community Detection Algorithm

We first describe the key idea used in this step, before presenting the algorithm in detail. Observe that, for any two reads that overlap in the sites covered, we can infer whether they belong to the same or different strings quite accurately. The intuition for this is identical to the pair-wise classify step of the community detection algorithm in the previous chapter. We can then leverage this to classify all nodes in a box, as good, using a non-parametric spectral method as opposed to the parametrized Is-Good sub-routine used in

the previous chapter. However, as the graph G has a large diameter (of order $\sqrt{n})$), all nodes cannot be classified accurately at once using the spectral method. Moreover, applying the spectral method to G is also computationally impractical, as any clustering schemes will be super-linear in the number of nodes N_n, which renders them horribly slow on large instances that are typical in practice. Furthermore, even pre-processing the reads to construct the graph G is of order N_n^2, which makes it infeasible on large instances. A direct spectral method on G is also statistically sub-optimal as the fluctuations in density of reads in space is variable, i.e., the point process representation is not homogeneous. In many problem instances, the density of reads can vary across space, just due to the randomness in the read generation process and spectral methods are sensitive to these variations. For instance, in Figure 6.1, we see an example where the density of reads captured by the spectral algorithm is highly imbalanced across the three strings due to the fluctuations of nodes in space. Hence, one would need an additional 'constraint' in the reconstruction algorithm to enforce the spatial distribution of reads across all communities to be identical. This is achieved by our Euclidean community detection algorithm, as we apply clustering on small windows of space. The partitioning based on spatial locality automatically ensures that the spatial distribution of the estimated communities are all roughly identical. The intuition for this comes from the fact that the reads will be roughly uniformly distributed within a box as a box is 'small' in size. As described in the previous chapter, we make the boxes overlapping, and hence a single read will be

182
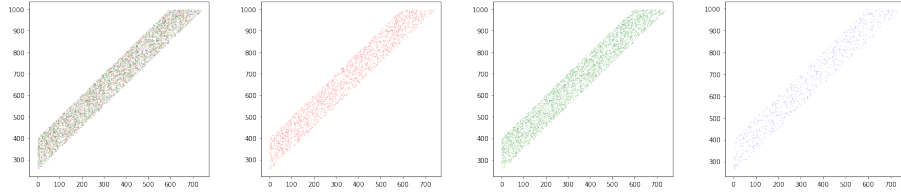
Figure 6.1: An instance with three strings and length of string being 1000, when spectral clustering on G fails. The coverage is 10, the read lengths $r = 2$ and $R = 250$. The error probability $p = 0.05$. The picture on the left is the union of the three pictures on the right. The three colored pictures are the spatial locations of the recovered communities by the spectral algorithm applied on G. The density of recovered blue estimates is 0.0995, while that of red is 0.33 and of green is 0.57. The total overlap achieved by the spectral method is 0.1. Note that in the ground truth, all three colors are equal in intensity, which is not captured by the spectral method. However, our algorithm predicts roughly equal sized communities and achieves an overlap of 0.98 and runs 4 times faster.

present in multiple boxes. This further boosts statistical accuracy of clustering by embedding natural *error-correction* scheme, as a single read gets estimated multiple times, each of which gives new 'evidence' for the community label of a node. From a computational complexity viewpoint, partitioning the set of nodes and clustering smaller instances dramatically reduces run-time as most typical clustering algorithms are super-linear in the number of data points and hence reducing the sizes of the graphs to be clustered has a significant impact. Thus, our algorithm is both computationally feasible on large instances and is statistically superior compared to standard graph clustering algorithms directly applied on G.

Pseudo Code

We set some notations that will be helpful in describing the algorithm, before going into the pseudo code. The algorithm has hyper-parameters $A, B, \text{iter}, M \in \mathbb{N}$ and $\alpha \in [0, 1]$. For $x, y \in [\lceil \frac{\tilde{n}}{A} \rceil]$, we denote by $B_{x,y} \subset [n]^2$ as $B_{x,y} := [Ax, \min(Ax + B, \tilde{n})] \times [Ay, \min(Ay + B, \tilde{n})]$, the box indexed by (x, y). Thus, the parameters A and B dictate how large a box is and how many boxes cover a single read. In the course of the algorithm, we maintain a dictionary of lists $\mathcal{C}$, where for each node $u \in [N_n]$, $\mathcal{C}[u]$ is a list of community estimate of node u. Each node has more than one estimate as it belongs to more than one box. The estimates from clustering in each box is added as 'evidence' of the community estimate for the node. Having multiple estimates for a node helps in combating clustering errors in certain boxes.

We now describe the algorithm in detail. The first step consists of partitioning the space $[n]^2$ into multiple overlapping boxes as shown in Figure 6.2. The hyper-parameters A and B allow one to tune both the size of a box, and also the number of boxes that will cover a given location of $[n]^2$. In each box indexed by (x, y) for $x, y \in [\lceil \frac{\tilde{n}}{A} \rceil]$, we identify the nodes of G having their spatial label in that box which we denote by $H_{x,y}$. If the number of nodes in $H_{x,y}$ is small, i.e., smaller than M, a hyper-parameter, then we do not attempt to cluster the nodes in this box. We set a minimum size for otherwise the output of the clustering will be very noisy and non-informative.Conversely, if more than α fraction of nodes in $H_{x,y}$ have at-least one community estimate, then again we do not cluster $H_{x,y}$. The reason for doing so and setting $\alpha < 1$ is

184
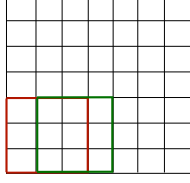
Figure 6.2: This is an example with $\tilde{n} = 7$. The parameters $A = 1$ and $B = 3$. The red and green boxes are examples of $B_{x,y}$.

to decrease the running-time by ensuring we only perform the clustering step when there are sufficiently new unexplored nodes. In each box (x, y), such that the number of nodes in $H_{x,y}$ is at-least M and at-most α fraction of them have prior estimates, we apply a fast and simple local clustering algorithm, i.e., a standard k-means algorithm [35] on the adjacency matrix of $H_{x,y}$. We then iteratively improve this initial clustering guess of the k-means algorithm by a combination of a linear transformation and a non-linearity. Once the nodes of $H_{x,y}$ are clustered, we then append this result to the dictionary of lists $\mathcal{C}$, after appropriately synchronizing the estimates in each node. We finally assign a single community estimate for each node based on the majority in the list of estimates in $\mathcal{C}$.

Local Clustering Step

This step is described in Algorithm 5. Here we follow a two step-procedure. In the first step, we get an approximate clustering of the graph H by applying standard k-means algorithm on the adjacency matrix H. We then 'one hot encode' this clustering result. More formally, if q denotes the

185

Algorithm 4 Main Routine

1: **procedure** MAIN$(G, k, (X_i)_{i \in [1, N_n]}, A, B, \text{iter}, M)$ ▷ $A, B, \text{iter}, M, \alpha$ are hyper-parameters.

2:     Tessellate $[n]^2$. For $x, y \in \left[\left\lceil \frac{n}{A} \right\rceil\right]$, let $B_{x,y} = [Ax, \min(Ax + B, n)] \times [Ay, \min(Ay + B, n)]$.

3:     **for** $1 \leq x \leq \left[\left\lceil \frac{n}{A} \right\rceil\right]$ **do**

4:         **for** $1 \leq x \leq \left[\left\lceil \frac{n}{A} \right\rceil\right]$ **do**

5:             $H_{x,y}$ - the subgraph of G with nodes spatial labels in $B_{x,y}$

6:             **if** Number of nodes in $H_{x,y} \geq M$ and fraction of nodes of $H_{x,y}$ with no prior estimate is larger than $1 - \alpha$ **then**

7:                 $\mathbf{e} \leftarrow$ Local-Cluster$(H_{x,y}, k, \text{iter})$

8:                 Update-Community-Estimates$(\mathcal{C}, \mathbf{e})$

9:             **end if**

10:         **end for**

11:     **end for**

12: **return** Reduce-by-Majority$(\mathcal{C})$

13: **end procedure**

---

number of nodes of H, then the one hot encoding result is a matrix $\mathbf{e}_0$ of size $q \times k$. Each entry of matrix $\mathbf{e}_0$ is either 0 or 1, with the entry in the ith row and jth column being 1 implies that the ith node is classified as belonging to community j. Thus, each row of the matrix $\mathbf{e}_0$ contains exactly one 1 and the rest of the entries are all 0. We then run a 'clean-up' procedure by iteratively updating the estimate as follows -

$$\mathbf{e}_{t+1} = \mathcal{T}(H \mathbf{e}_t). \tag{6.3}$$

The function $\mathcal{T}$ is applied row-wise, where for a matrix A, it sets the ith row and jth column of $\mathcal{T}(A)$ to 1 if $j = \arg\max A[i]$, else the ith row and jth column of $\mathcal{T}(A)$ is set to 0. If a row has more than one column where the maximum is attained, the first column where the maximum occurs gets a

---
Algorithm 5 Small Graph Clustering

1: **procedure** LOCAL-CLUSTER(H, k, iter)
2:     $\tilde{\mathbf{e}} := \text{k-means-cluster}(H, k)$
3:     $\mathbf{e}_0 = \text{One-Hot-Encoding}(\tilde{\mathbf{e}})$.
4:     **for** $1 \leq i \leq$ iter **do**
5:         $\mathbf{e}_i = \arg\max H\mathbf{e}_{i-1}$ ▷ Argmax is row wise on the product
6:     **end for**
7: **return** $\mathbf{e}_{\text{iter}}$.
8: **end procedure**
---

value 1 and the other columns get a value of 0. Hence the dimension of matrix A and $\mathcal{T}(A)$ are the same. Furthermore, for any matrix A, the matrix $\mathcal{T}(A)$ is such that its entries are either 0 or 1, with each row having exactly one entry valued at 1.

The iterative update is based on the following intuition. Let the clustering encoded as the matrix $\mathbf{e}_t$ for some $t \in \mathbb{N}$, and consider a tagged node $u \in [n]$. The new updated value of the community label of node u is then the 'most-likely' label *given* the estimates for the rest of the nodes are kept fixed. More precisely, the 'weight' that node u is in a community $l \in [k]$ is the sum of weights along all edges connecting u to $v \in [n]$ in the graph G such that the estimate of node v is l. The new community label of node u is that which has the highest weight. Performing this operation simultaneously for all nodes, one obtains the representation in Equation (6.3). It is clear that, for the iterative update to work well, the initial estimate $\mathbf{e}_0$ must be 'good enough', as a random community assignment to all nodes is a distributional fixed point to the recursion in Equation (6.3). In principle, one can get somewhat better

initial guess for $\mathbf{e}_0$ by applying the k-means algorithm to the eigenvectors of H, but the marginal gains in statistical accuracy does not warrant the enormous increase in computation needed to perform such a spectral clustering.

The clean-up method, at first glance, seems to bear similarities to other dynamical algorithms such as expectation maximization, Belief Propagation (BP) and tensor factorization based methods of [193]. However, unlike BP, we do not iterate the beliefs or probabilities of a node belonging to various communities, instead we make a hard assignment at each update step. The reason for doing so, is one it is *non-parametric* and does not rely on any explicit generative models of the data. Secondly, the graph G is not tree-like and has a lot of triangles and loops due to the spatial embedding. Hence, we cannot just keep track of the node marginals, but instead need the whole joint distribution, and is hence not tractable. In light of these structural observations, we use the heuristic recursion defined in Equation (6.3).

Synchronization Step

The main routine in Algorithm 4 considers the boxes sequentially and performs local clustering step. Once the local clustering is performed, a key component is to *synchronize* the estimates of the current box with the estimates of the boxes that are already clustered. A synchronization is essential since the problem is permutation invariant to the labels. Formally, the statistical distribution of the data remains unchanged if the true underlying labels of the $[k]$ strings are permuted. Hence, the best hope for any recovery algo-

188

Algorithm 6 Synchronization Step
___
1: **procedure** SYNCHRONIZE($\mathcal{C}, \mathbf{e}$)
2:     max-weight $\leftarrow 0$
3:     max-perm $\leftarrow \pi$.
4:     **for** All permutations π of $[k]$ **do**
5:         local-weight $\leftarrow 0$
6:         **for** All nodes in e **do**
7:             local-weight $\leftarrow$ local-weight $+$fraction($\mathcal{C}$[node]$, \pi(\mathbf{e}$[node]).
8:         **end for**
9:         **if** local-weight $>$ max-weight **then**
10:             max-weight $\leftarrow$ local-weight
11:             max-perm $\leftarrow \pi$.
12:         **end if**
13:     **end for**
14:     **for** nodes in $\mathbf{e}$ **do**
15:         Append max-perm($\mathbf{e}$[node]) to $\mathcal{C}$[node].
16:     **end for**
17: **return** $\mathcal{C}$.
18: **end procedure**
___

rithm is to recover the k strings upto a permutation of labels. Thus, if any clustering algorithm is run on two different subsets of nodes, the corresponding estimates need to be synchronized to produce a clustering of the nodes in the union of the sets. We do this clustering in Line 8 of main routine Algorithm 4 by invoking sub-routine 6.

In sub-routine 6, we decide on how to permute the label output of the local-clustering estimate of $H_{x,y}$, that best 'synchronizes' with the estimates of other nodes of G at that instance. Observe that at the instant of synchronizing the output of $H_{x,y}$, the rest of the nodes of G have either none or multiple estimates per node. There is possibility of more than one estimate per node as

Algorithm 7 Reduce by Majority

1: **procedure** REDUCE-BY-MAJORITY($\mathcal{C}$)
2:     **for** node in $\mathcal{C}$'s keys **do**
3:         $\mathcal{C}[\text{node}] \leftarrow \text{Majority}(\mathcal{C}[\text{node}])$.
4:     **end for**
5: **return** $\mathcal{C}$
6: **end procedure**

---

a node is present in multiple boxes, each of which adds an estimate as 'evidence' for a node's cluster. We select a permutation of the labels by sequentially going over all permutations of $[k]$ and selecting the one that has the highest 'synchronization-weight'. More formally, let $\mathcal{N}_{x,y} \subset [n]$ denote the indices of the nodes in $H_{x,y}$ and for all $u \in \mathcal{N}_{x,y}$, denote by $\sigma[u] \in [k]$ be the estimate of the local-clustering on $H_{x,y}$. The synchronization weight for a permutation π of $[k]$, is defined by

$$W_\pi := \sum_{i \in \mathcal{N}_{x,y}} \mathbf{1}_{\mathcal{C}[i] \neq \text{empty}} \frac{\sum_{j \in \mathcal{C}[i]} \mathbf{1}_{\pi(\sigma[i]) = \mathcal{C}[i][j]}}{\sum_{j \in \mathcal{C}[i]} 1}.$$

In words, we go over all nodes in $H_{x,y}$ that have at-least one prior estimate and sum the fraction of the previous estimate equaling the label assigned by the local-clustering $H_{x,y}$ after applying the permutation π to the local clustering's output. Among all permutations, we select a π^* having the highest synchronization weight breaking ties arbitrarily. After doing so, we append to each node u of $H_{x,y}$, the label $\pi^*(\sigma(u))$ to the list $\mathcal{C}[u]$. The key feature above is to consider the *fraction*, which is a proxy for the 'belief' of the community label of a node, rather than just a count, as the counts across different nodes can be very skewed due to the order in which the boxes are clustered and

synchronized.

Computational Complexity

In this section, we discuss the computational complexity of implementing our algorithm, and the role played by the various hyper-parameters in this regard. A naive implementation of our algorithm will incur a cost of order n^2, just to construct the graph G from the reads. This step itself will be infeasible in practical scenarios where the number of reads will be in the order of millions. However, our algorithm only needs the subgraphs $H_{x,y}$ and not the full graph G. To do so, we pre-process the reads and create a hash-map, where for each location in $[n]^2$, we store the list of reads that has its spatial label in that location. This takes a one pass through the list of reads, i.e., has complexity of order n and storage complexity of order n. Now, to create the adjacency matrix $H_{x,y}$, we only take a time quadratic in the number of nodes in $H_{x,y}$. The synchronization step takes time complexity of the order of number of nodes in $H_{x,y}$ times the number of distinct permutations of $[k]$.

Our algorithm also has an inherent trade-off in computational complexity and statistical accuracy by varying the hyper-parameters. For instance, if we decrease A, while keeping B fixed, then we increase the number of boxes, thereby increasing the computational time. However, the statistical accuracy will improve since each node will now be present in many boxes and hence the error-correction scheme will not perform worse off. Similarly, increasing the parameter M can reduce run-time by considering only a fewer boxes to

perform local clustering, while potentially decreasing statistical accuracy, as each node will have fewer evidence.

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.05	94.537	559.05	5.06	9.55	96.34
5	0.1	82.14	1209.17	5.09	16.23	87.57
	0.2	41.71	2317.02	5.07	6.96	58
	0.05	99.66	739.65	8.01	0.23	99.82
8	0.1	96.44	1561.46	7.4	8.28	97.56
	0.2	57.97	3597.56	7.35	12.49	69.38
	0.05	99.88	930.61	9.93	0.11	99.94
10	0.1	98.87	1867.85	10.06	5.29	99.24
	0.2	69.41	4298	10.47	17.59	77.45

Table 6.1: The simulation results for ComHapDet diploid biallelic case.

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.05					
5	0.1					
	0.2					
	0.05					
8	0.1					
	0.2					
	0.05					
10	0.1					
	0.2					

Table 6.2: The simulation results of AltHap for diploid biallelic case.

6.4 Experimental Evaluation

We benchmark the performance of our algorithm through simulations, in both the diploid biallelic case as well as the more challenging polyploid polyallelic case. Since, ground truth is known in simulations, we use CPR,

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.002	91.72	293.2	53.63	10.344	96.28
5	0.01	93.85	349.8	48	2.57	97.7
	0.05	78.2	1734.4	50.46	18.55	89.13
	0.002	98.6	97	76.7	0.88	99.5
7	0.01	93.78	662.1	81.25	10.794	96.95
	0.05	97.11	1504.7	75.52	1.571	98.9
	0.002	99.75	93.7	137.5	0.168	99.91
10	0.01	99.67	413.1	135.9	0.21	99.89
	0.05	99.44	2021.9	139.78	0.27	99.77
	0.002	99.91	124.6	300.35	0.11	99.97
15	0.01	99.88	611.1	307.88	0.07	99.95
	0.05	99.86	2981.5	297.19	0.15	99.95

Table 6.3: Simulated Triploid Tetraallelic Data with ComHapDet.

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.002	73.97	1479.7	124.59	22.6	83.4
5	0.01	76.73	1501.1	173.89	13.8	85.61
	0.05	65.33	3165.9	190.95	17.79	77.91
	0.002	88.95	687	295.22	13.97	92.97
7	0.01	88.69	966.2	289.75	17.5	92.44
	0.05	80.13	2887.4	332.1	20.27	86.31
	0.002	83.67	1215.4	593.19	20.65	88.42
10	0.01	92.72	1029.1	592.74	14.59	95.36
	0.05	92.73	1029.1	592.44	14.59	95.36
	0.002	89.89	1725	708.5	16.07	94
15	0.01	95.96	1628.6	781	9.82	97.58
	0.05	87.43	6721.3	713.3	20.36	92.09

Table 6.4: Simulated Triploid Tetraallelic Data.

MEC and M-CPR as the primary performance benchmarks. We compare the performance of our algorithm against AltHap [193], the state of the art method for haplotype phasing in both diploid and polyploid settings. In the diploid

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.002	39.2	2996.1	76.57	13.96	71.55
5	0.01	49.76	2717.3	86.96	17.08	78.10
	0.05	30.16	4642.7	77.99	8.57	67.325
	0.002	79.97	1316.25	143.48	20.27	91.8
7	0.01	79.09	1640.0	118.52	17.84	91.8
	0.05	68.34	3722.8	129.66	13.98	87.29
	0.002	98.86	193.1	253.32	1.42	99.64
10	0.01	99.17	32.63	261.81	0.41	99.76
	0.05	98.2	2727.7	238.56	0.64	99.51
	0.002	99.75	182.7	487.02	0.22	99.93
15	0.01	99.75	806.5	482.74	0.18	99.94
	0.05	99.0	4101.4	523.78	298.88	99.65

Table 6.5: Simulated Tetraploid Tetraallelic Data.

Coverage	Error Rate	CPR	MEC	t(sec)	σ(CPR)	M-CPR
	0.002	61.2	2340.6	308.98	32.82	79.13
5	0.01	56.19	2892.6	323.38	21.35	77.62
	0.05	37.88	5431.4	422.92	23.4	66.01
	0.002	76.08	1388.6	521.36	20.81	87.49
7	0.01	79.86	1812.8	515.78	20.45	88.05
	0.05	83.59	3481.9	503.13	20.23	91.97
	0.002	71.92	1979.7	594.3	15.5	85.58
10	0.01	85.44	1779.4	585	18.53	92.10
	0.05	78.55	5331.4	667.49	15.55	89.65
	0.002	85.21	2614.6	684.45	18.39	92.01
15	0.01	83.53	3973.7	684.13	17.41	92.61
	0.05	95.13	6397.6	682.51	14.47	97.38

Table 6.6: Simulated Tetraploid Tetraallelic Data.

biallelic case, we use the synthetic data from [177] which is often used as a benchmark to compare methods for haplotype assembly. We report the results in Tables 6.1 and 6.2. In the polyploid case, we benchmark the performance

of our algorithm using the synthetic data from [193]. We report the results in Tables 6.4, 6.3 for the triploid case and Tables 6.6, 6.5 for the tetraploid cases. **??** for the cases of triploid, tetraploid and hexaploid respectively. In both the triploid and tetraploid cases, we considered the tetraallelic setting, i.e., the alphabet cardinality is 4. The average read length of a single end of the paired-end read is 2. The average insert size between the paired end reads was 200. In each polyploid case, we test our algorithm on 10 different problem instances, where in each instance, a haplotype sequence of length 1000 was phased. We use the same method and the publicly available code from [193] to generate the synthetic data for the various instances.

The CPR and M-CPR are averages, reported after multiplying by 100, for ease of comparison. We compare the performance as the problem parameters, namely the ploidy and alphabet size as well as the measurement parameters, namely the coverage, average read length and error rates are varied. In each case, the hyper-parameters was set to $A = 15$ and $B = 4$, $\alpha = 0.95$, for all polyploid cases and used the parameters $A = 20$, $B = 2$ and $\alpha = 0.7$ for all the diploid biallelic cases. The column $\sigma(\text{CPR})$ denotes the standard deviation of CPR after multiplying it by 100. As expected, the problem gets harder, i.e., both the run times increase, as well as CPR reduces, as the ploidy increases, or if the coverage reduces. We implemented our algorithm in Python, and both the simulations, as well as the experimental evaluations were conducted on a single core Intel I5 Processor with 2.3Ghz processor and 8 GB 2133 MHz LPDDR3 RAM.

6.4.1 Results and Discussions

We observe from the simulations that our method performs very well in the polyploid polyallelic scenario, significantly outperforming the previous state of the art method of [193]. In the diploid biallelic case, we see that our method is competitive with the state of the art, with both methods achieving near perfect reconstruction at reasonable run times. In terms of complexity, we notice that our method is fast, even in the polyploid polyallelic case, and scales gracefully, both with ploidy and coverage. Importantly, we notice that the runtime of our algorithm also scales very gracefully with problem size, and in particular, does not have exponential dependence on any problem parameters. Indeed, with increasing coverage, we can tune the α parameter to both achieve superior CPR, while retaining very reasonable run-times. Even with increasing read-lengths, our problems run-time essentially remains unchanged, which makes it attractive for high coverage scenarios.

6.5 Conclusions

In this chapter, we proposed a novel representation of the Haplotype phasing problem as an instance of Euclidean community detection. Motivated by the algorithmic study in Chapter 5, we propose a novel algorithm. We see in our experimental studies that the proposed algorithm performs very well in simulations and is comparable to the state of art methods for Haplotype Phasing, both in terms of statistical accuracy and runtime performance.

Appendices

Appendix A

Proofs from Chapter 2

A.1 Proof of Theorem 1

Proof. We prove this by contradiction. Assume that ϕ_t is in stationary regime and that $\lambda > \frac{Cl(T)}{\ln(2)La}$. We use the Miyazawa's Rate-Conservation Principle or Law (RCL) (e.g. [40], 1.3.3) to set-up a system of equations and identify a contradiction. Applying the RCL to the stochastic process $\phi_t(\mathbf{S})$ which counts the number of links yields,

$$\lambda |\mathbf{S}| = \lambda_d, \tag{A.1}$$

where λ_d is the intensity of the point-process on $\mathbb{R}$ corresponding to the epochs of a death-time. Since we assumed that ϕ_t is in stationary regime, the point process formed on the real line by the instants of a death is stationary with intensity $\lambda_d = \lambda |\mathbf{S}|$. Applying RCL to the total "work-load" in the network i.e. the total number of bits that each of the transmitters present are yet to send to their corresponding receivers, we get

$$\lambda |S| L = \mathbb{E}\left[\sum_{x \in \phi_0} R(x, \phi_0)\right], \tag{A.2}$$

where $R(x, \phi)$ is given in Equation (2.1). From the definition of Palm Probability of ϕ_0, we have that

$$\lambda |\mathbf{S}| L = \mathbb{E}^0_{\phi_0}\left[R(0, \phi_0)\right] \mathbb{E}[\phi_0(\mathbf{S})], \tag{A.3}$$

198

where $\mathbb{E}^0_{\phi_0}$ is the (spatial) Palm Probability of ϕ_0 and $\phi_0(\mathbf{S})$ is the random variable denoting the number of links in the network in steady-state. Note that from our assumption that ϕ_t is in stationary regime ensures the existence of the Palm Probability measure of the spatial point process ϕ_0. Applying rate-conservation to the stochastic process $\mathbf{I}_t = \sum_{x \in \phi_t} I(x, \phi_t)$, the sum interference seen at all receivers (which could possibly be ∞), we get

$$\lambda |\mathbf{S}| \mathbb{E}^\uparrow[\mathcal{I}] = \lambda_d \mathbb{E}^\downarrow[\mathcal{D}], \qquad (A.4)$$

with $\mathcal{I} = \mathbf{I}_{0_+} - \mathbf{I}_0$ and $\mathcal{D} = \mathbf{I}_0 - \mathbf{I}_{0_+}$. Here, $\mathbb{E}^\uparrow$ denotes the (time) Palm probability corresponding to the point process on $\mathbb{R}$ of birth instants and $\mathbb{E}^\downarrow$ denotes the (time) Palm probability of the point process on $\mathbb{R}$ corresponding to the instants of death. From Equation (A.1) we have

$$\mathbb{E}^\uparrow[\mathcal{I}] = \mathbb{E}^\downarrow[\mathcal{D}]. \qquad (A.5)$$

From the PASTA property and the fact that the births are uniform in $\mathbf{S}$, we have from Campbell's theorem that

$$\mathbb{E}^\uparrow[\mathcal{I}] = 2\mathbb{E}[\phi_0(\mathbf{S})]\frac{a}{|\mathbf{S}|}. \qquad (A.6)$$

Since the file-sizes at all transmitters are i.i.d. exponential with mean L, the point process on the real line corresponding to the death-instants admits as stochastic-intensity $\mathbf{R}_t = \frac{1}{L}\sum_{x \in \phi_t} R(x, \phi_t)$ with respect to the filtration $\mathcal{F}_t = \sigma(\phi_s : s \leq t)$, the sigma algebra corresponding to the locations. Hence, it then follows from Papangelou's theorem (e.g. [40], Theorem 1.9.2) that

$$\frac{d\mathbb{P}^\downarrow}{d\mathbb{P}}|_{\mathcal{F}_{0_-}} = \frac{\mathbf{R}_0}{\mathbb{E}[\mathbf{R}_0]}. \qquad (A.7)$$

Since the decrease in total interference (in state ϕ_{0_-}) is of magnitude $I(X, \phi_0)$ with probability $\frac{R(X,\phi_0)}{L\mathbf{R}_0}$ if $X \in \phi_{0_-}$, we get

$$\mathbb{E}^{\downarrow}[\mathcal{D}] = 2\mathbb{E}\left[\frac{\mathbf{R}_0}{\mathbb{E}[\mathbf{R}_0]} \sum_{x \in \phi_0} \frac{R(x, \phi_0)}{L\mathbf{R}_0} I(x, \phi_0)\right]$$
$$= 2\frac{\mathbb{E}[\sum_{x \in \phi_0} R(x, \phi_0) I(x, \phi_0)]}{L\mathbb{E}[\mathbf{R}_0]}$$
$$= 2\frac{\mathbb{E}^0_{\phi_0}[R(0, \phi_0) I(0, \phi_0)]}{L\mathbb{E}[\mathbf{R}_0]}\mathbb{E}[\phi_0(\mathbf{S})]. \qquad (A.8)$$

Now combining, Equations (A.8), (A.6) and (A.2), we get

$$a = \frac{\mathbb{E}^0_{\phi_0}[R(0, \phi_0) I(0, \phi_0)]}{L\lambda}. \qquad (A.9)$$

From Equation (2.1) and basic calculus, we have that $R(0, \phi_0) I(0, \phi_0) \leq \frac{Cl(T)}{\ln(2)}$ which is a deterministic bound that is true for *any* $\phi \in \mathbf{M}(\mathbf{S})$. Applying this inequality to Equation (A.9), we get the inequality that

$$\lambda \leq \frac{Cl(T)}{\ln(2)La}. \qquad (A.10)$$

Inequality (A.10) is a contradiction to our assumption that ϕ_t is in stationary regime and that $\lambda > \frac{Cl(T)}{\ln(2)La}$. $\qquad \square$

Appendix B

A.2 Proof of Theorem 3

For simplicity of the proof, we assume the link distance $T = 0$. The proof for arbitrary T follows with significantly more notation that obscures the essence of the proof. Thus, to keep the proof ideas simple, we first outline

the proof for the special case of $T = 0$ with remarks in between as to how the intermediate steps can generalize. At the end, we will give the complete construction of the coupling (which will be explained later) in the general case of T being arbitrary. This will then complete the proof in the general case as well.

Assume $T = 0$ for the time being. Thus, the dynamics is that of points arriving and exiting the network and the network at any point of time consists of a collection of points distributed in space $\mathbf{S}$. The high level idea of the proof is that we tesselate the space $\mathbf{S}$ and study another "upper-bound" Markov Chain living on a countable state-space which we analyze through fluid limit techniques. We then conclude about the ergodicity of ϕ_t which is a Markov Chain on the topological space $\mathbf{M}(\mathbf{S})$.

To define the upper-bound chain, we first tessellate the square $\mathbf{S}$ into cells where each cell is a square of length exactly ϵ. Since $\mathbf{S}$ is a torus, we assume without loss of generality that the origin is in the center of a cell. One can find a sequence of such tessellations with the side length of the cells going to 0. The tessellation for each valid $\epsilon > 0$ results in n_ϵ, a finite number of cells as $\mathbf{S}$ is compact. Index the cells by i and let A_i denote the subset of $\mathbf{S}$ corresponding to cell i and $a_i \in A_i$ denote its center. The cell containing the origin is indexed 0 i.e. $a_0 = 0$. For such an ϵ tessellation, we define a new path-loss function $l_\epsilon(x, y)$ where $l_\epsilon(x, y) = l_\epsilon(a_i, a_j)$ for all $x \in A_i$ and $y \in A_j$

201

and

$$l_\epsilon(a_i, a_j) = \sup\{l(||b_i - b_j||) : ||b_i - a_i||, ||b_j - a_j|| \in \{0, \epsilon\}\}.$$

Note that the function l_ϵ satisfies

$$\sum_i l_\epsilon(a_i, a_j) = \sum_i l_\epsilon(a_i, 0) = \frac{1}{\epsilon^2} \int_{x \in \mathbf{S}} l_\epsilon(||x||) dx, \qquad (A.11)$$

since $\mathbf{S}$ is a square torus and each cell A_i is a square of side-length ϵ,

The upper bound Markov-Chain is denoted as $\phi_t^{(\epsilon)}$ which takes value in the space $\mathbf{M(S)}$. This chain has the exact same dynamics as described in Equation (2.4) *except* that the interference comes from $l_\epsilon(.,.)$ instead of from $l(\cdot)$,

Lemma 82. For all time t, the point-process $\phi_t^{(\epsilon)}$ stochastically dominates ϕ_t. This implies that if $\phi_t^{(\epsilon)}$ is stable for a particular λ, then so is ϕ_t for that value of λ.

Proof. We have from the monotonicity of $l(\cdot)$, $l_\epsilon(x, y) \geq l(x, y) = l(||x - y||)$ for each $x, y \in \mathbf{S}$. Thus, for each $x \in \mathbf{S}$ and each $\phi \in \mathbf{M(S)}$, $I_\epsilon(x, \phi) \geq I(x, \phi)$ and subsequently $R_\epsilon(x, \phi) \leq R(x, \phi)$ as $R(x, \phi)$ is a decreasing function of $I(x, \phi)$. Therefore the point process $\phi_t^{(\epsilon)}$ *stochastically dominates* the point process ϕ_t. This follows from the fact that for any $\phi \in \mathbf{M(S)}$, we have that the birth rate $\lambda|S|$ is the same for both process, whereas the death-rate of each point of $x \in \phi$ satisfies $\frac{1}{L} R_\epsilon(x, \phi) \leq \frac{1}{L} R(x, \phi)$. Also form Equation (2.1), if $\phi_1 \subseteq \phi_2$, then for each $x \in \phi_1 \cap \phi_2$, $R(x, \phi_1) \geq R(x, \phi_2)$. Hence one can construct a

coupling of the process ϕ_t and $\phi_t^{(\epsilon)}$ such that $\phi_t \subseteq \phi_t^{(\epsilon)} \; \forall t$, i.e. a point is alive in ϕ_t only if it is also alive in $\phi_t^{(\epsilon)}$. Therefore, if $\phi_t^{(\epsilon)}$ is ergodic for a given λ, then ϕ_t is also ergodic for that arrival rate λ. $\qquad\qquad\qquad\qquad\square$

Define $\mathbf{X}^{(\epsilon)}(t) = \{\phi_t^{(\epsilon)}(A_i)\}_{i=1}^{n_\epsilon}$ as the n_ϵ dimensional vector taking values in $\mathbb{N}^{n_\epsilon}$. It is easy to see that $\mathbf{X}^{(\epsilon)}(t)$ is a Markov-Chain since the path-loss function $l_\epsilon(x,y)$ does not distinguish between two different locations of space inside a cell. It is also evident that if $\mathbf{X}^{(\epsilon)}(t)$ is ergodic, then $\phi_t^{(\epsilon)}$ is ergodic since $\lim_{t \to \infty} \mathbb{P}[\phi_t^{(\epsilon)}(\mathbf{S}) < \infty] = \lim_{t \to \infty} \mathbb{P}[||\mathbf{X}^{(\epsilon)}(t)||_1 < \infty] = 1$. The second equality follows from the fact that $\mathbf{X}^{(\epsilon)}(t)$ is a finite-dimensional ergodic Markov chain on $\mathbb{N}^{n_\epsilon}$. Hence, a sufficient condition for stability of ϕ_t is a condition for the Markov Chain $\mathbf{X}^{(\epsilon)}(t)$ to be ergodic.

We show in Theorem 83 that $\mathbf{X}^{(\epsilon)}(t)$ (and hence $\phi_t^{(\epsilon)}$) is ergodic if

$$\lambda < \frac{C}{L \ln(2) \int_{x \in \mathbf{S}} l^\epsilon(x, 0) dx}, \qquad\qquad (A.12)$$

which will actually conclude the proof of Theorem 3. This can be seen as follows. Since the point process ϕ_t^ϵ stochastically dominates ϕ_t, we can optimize the stability region in Equation (A.12) by choosing the best ϵ. As the function $r \to l(r)$ is monotone, $l_\epsilon(x, 0)$ is monotone increasing in ϵ for each $x \in \mathbf{S}$ and hence we want to have ϵ as small as possible. Furthermore, the function $r \to l(r)$ has only a countable set of discontinuity points (as it is bounded non-increasing), we have that as ϵ goes to 0, $l_\epsilon(x, 0)$ converges to $l(x, 0)$ for almost-every $x \in \mathbf{S}$. Hence, $\lim_{\epsilon \to 0} \int_{x \in \mathbf{S}} l_\epsilon(||x||) dx = \int_{x \in \mathbf{S}} l(||x||) dx$ from the

Monotone Convergence theorem. Therefore, if $\mathbf{X}^{(\epsilon)}(t)$ is ergodic under condition in Equation (A.12), then ϕ_t will be ergodic under the condition

$$\lambda < \limsup_{\epsilon \to 0} \frac{C}{L \ln(2) \int_{x \in \mathbf{S}} l^\epsilon(x, 0) dx} = \frac{C}{L \ln(2) \int_{x \in \mathbf{S}} l(x, 0) dx}, \qquad \text{(A.13)}$$

which will conclude the proof of Theorem 3.

Theorem 83. $\mathbf{X}^{(\epsilon)}(t)$ is ergodic under the condition in Equation (A.12).

We remark that, even in the general case of $T > 0$, the same theorem statement will hold for a slightly modified version of $\mathbf{X}^{(\epsilon)}(t)$ which we will construct later. Thus, if Theorem 83 is established for arbitrary T, the proof of the main theorem will be complete by similar reasoning in the previous paragraph.

Proof. We can write the following evolution for the vector $\mathbf{X}^{(\epsilon)}(t)$ which we refer to as $\mathbf{X}(t)$ in the sequel for convenience as

$X_i \to X_i + 1$ at rate $\lambda \epsilon^2$

$X_i \to X_i - 1$ at rate

$$X_i \log_2 \left(1 + \frac{1}{N_0 + \sum_{j=1}^{n_\epsilon} (X_j - \mathbf{1}(j = i)) l_\epsilon(a_i, a_j)} \right).$$

We note that generalizing this dynamics to the case when $T > 0$ is slightly different alebit the same principles and we outline it at the end of the proof.

Under condition in Equation (A.12), we show the following drift argument to hold which will conclude the proof.

Theorem 84. [322] Let $\mathbf{X}(t)$ be a Markov Chain taking values in a countable state space $\mathcal{S}$. Assume there exists a function $L : \mathcal{S} \to \mathbb{R}_+$ and constants $A < \infty$, $\epsilon > 0$ and an integrable stopping time $\hat{\tau} > 0$ such that for all $x \in \mathcal{S}$:

$$L(x) > A \implies \mathbb{E}_x L(\mathbf{X}(\hat{\tau})) \leq L(x) - \epsilon \mathbb{E}_x(\hat{\tau}). \tag{A.14}$$

If in addition the set $\{x : L(x) \leq A\}$ is finite and
$\mathbb{E}_x L(\mathbf{X}(1)) < \infty$ for all $x \in \mathcal{S}$, then $\mathbf{X}(t)$ is ergodic.

We will show that the above theorem is satisfied with the Lyapunov function $L(x) = ||x||_\infty$ and

$$
\begin{aligned}
\hat{\tau} &= L(\mathbf{X}(0)) \left(\frac{C}{L \ln(2) \sum_{k=0}^{n_\epsilon - 1} l_\epsilon(a_k, 0)} - \lambda \epsilon^2 \right)^{-1} \\
&:= L(\mathbf{X}(0))\tau,
\end{aligned} \tag{A.15}
$$

a deterministic finite stopping-time. We will use the notation that $||x||_\infty = |x|$ which is also equal to $L(x)$.

To establish the drift condition, we pass to the fluid-limit. A fluid limit of the Markov-Process $\mathbf{X}(t)$ is denoted by $x(t)$ which is a n_ϵ dimensional vector. $x(t)$ is defined as a fluid limit if there exists non-decreasing Lipschitz continuous function $\{D_i(t)\}_{i=0}^{n_\epsilon - 1}$ such that

$$x_i(t) = x_i(0) + \lambda \epsilon^2 t - D_i(t),$$

where the derivative of $D_i(t)$ satisfies $\dot{D}_i(t) = \frac{Cx_i(t)}{L\ln(2)\sum_{j=0}^{n_\epsilon-1} x_j(t)l_\epsilon(a_i,a_j)}$, or equivalently, the fluid limit $x(t)$ satisfies the following set of differential equations. If $||x(t)||_\infty > 0$,

$$\frac{d}{dt}x_i(t) = \lambda\epsilon^2 - \frac{Cx_i(t)}{L\ln(2)\sum_{k=0}^{n_\epsilon-1} x_k(t)l_\epsilon(a_i,a_k)} \tag{A.16}$$

and if $||x(t)||_\infty = 0$,

$$\frac{d}{dt}x_i(t) = 0.$$

For $y \in \mathbb{R}^{n_\epsilon}$, denote by $S(y)$ the set of fluid functions $x(t)$ such that $x(0) = y$. The following theorem establishes that the above fluid equation is indeed obtained through an appropriate space and time scaling. It also establishes as a corollary that $S(y)$ is non-empty for any $y \in \mathbb{R}^{n_\epsilon}$.

Theorem 85. Consider a sequence of deterministic initial conditions $\{X^{(k)}(0)\}_{k\geq 1}$ for the Markov Chain $\mathbf{X}(t)$ and a sequence of positive integers $\{z_k\}_{k\geq 1}$ with $\lim_{k\to\infty} z_k = \infty$ such that the limit $\lim_{k\to\infty} z_k^{-1} X^{(k)}(0) = x(0)$ exists. Then for all $s > 0$ and all $\delta > 0$, the following convergence takes place

$$\lim_{k\to\infty} \mathbb{P}\left(\inf_{f\in S(x(0))} \sup_{t\in[0,s]} |z_k^{-1}\mathbf{X}^{(k)}(z_k t) - x(t)| > \delta\right) = 0.$$

This proof is standard and is postponed later on in the appendix.

From the description of the dynamics, if we have $L(x(t)) = 0$, then $x_i(t) = 0$ for all i. Since $x(t)$ is a finite-dimensional vector, there exists at-

least one coordinate $i^*(t)$ such that $x_{i^*(t)}(t) = L(x(t))$. Then one can write

$$\frac{d}{dt}L(x(t)) = \lambda\epsilon^2 - \frac{CL(x(t))}{L\ln(2)\sum_{k=0}^{n_\epsilon-1}x_k(t)l_\epsilon(a_{i^*(t)}, a_k)}$$

$$\leq \lambda\epsilon^2 - \frac{C}{L\ln(2)\sum_{k=0}^{n_\epsilon-1}l_\epsilon(a_k, 0)}, \tag{A.17}$$

where the second inequality comes by the fact that $x_k(t) \leq L(x(t))$ and the symmetry of the torus as given in Equation (A.11). From Equation (A.17), we see that under the condition given in (A.12), $L(x(s)) = 0$ for all $s \geq \tau$ whenever $L(x(0)) = 1$, where $\tau = \left(\frac{C}{L\ln(2)\sum_{k=0}^{n_\epsilon-1}l(a_k,0)} - \lambda\epsilon^2\right)^{-1}$, a deterministic time as defined in Equation (A.15).

We remark that inequality (A.17) will be identical even in the case of arbitrary link distance T and hence, the rest of the proof ingredients are the same for both when $T = 0$ and $T > 0$.

Lemma 86. If condition in Equation (A.12) holds, then

$$\lim_{L(x)\to\infty}\frac{1}{L(x)}\mathbb{E}_x[|\mathbf{X}(L(x)\tau)|] = 0. \tag{A.18}$$

where τ is defined in Equation (A.15)

Proof. The first observation is that the family of random variables $\left\{\frac{|\mathbf{X}_x(|x|t)|}{|x|}\right\}_{x\in\mathbb{N}^{n_\epsilon}\setminus\{0\}}$ is uniformly integrable. Indeed, let $\{A_i(\cdot)\}_{i=0}^{n_\epsilon-1}$ be i.i.d. unit rate PPP denoting the arrivals into cell i. Then

$$X_i(t) \leq X_i(0) + A_i(\lambda\epsilon^2 t). \tag{A.19}$$

207

Thus, for $\mathbf{X}(0) = x$,

$$\frac{X_i(|x|\tau)}{|x|} \leq \frac{x_i}{|x|} + \frac{A_i(\lambda\epsilon^2|x|\tau)}{|x|}. \tag{A.20}$$

We have that $\frac{x_i}{|x|} \leq 1$ and the mean of $\frac{A_i(\lambda\epsilon^2|x|\tau)}{|x|}$ equal to $\lambda\epsilon^2\tau$. The variance of $\frac{A_i(\lambda\epsilon^2|x|\tau)}{|x|}$ is $\frac{\lambda\epsilon^2\tau}{|x|} \leq \lambda\epsilon^2\tau$ for all $x \in \mathbb{N}^{n_\epsilon} \setminus \{0\}$. As the variance is uniformly bounded, the random variables $\left\{\frac{X_i(|x|\tau)}{|x|}\right\}_{x \in \mathbf{N}^{N_\epsilon}\setminus\{0\}}$ are uniformly integrable. In addition, $\frac{|\mathbf{X}(|x|\tau)|}{|x|} \leq \sum_i \frac{X_i(|x|\tau)}{|x|}$, gives that $\left\{\frac{|\mathbf{X}_x(|x|\tau)|}{|x|}\right\}_{x \in \mathbf{N}^{N_\epsilon}\setminus\{0\}}$ is uniformly integrable since it is bounded above by a finite sum of random variables belonging to uniformly integrable families.

Let x_k be any sequence of initial conditions such that $|x_k| \to \infty$. This implies that $a_k = \mathbf{X}^{(k)}(0)/|x_k| = x_k/|x_k|$ with $a_k \in [-1, 1]^{n_\epsilon}$ for all k. Since the cube $[-1, 1]^{n_\epsilon}$ is compact, there is a convergent sub-sequence i.e. $\frac{\mathbf{X}^{k(l)}(0)}{|x_{k(l)}|} \to x(0)$ with $|x(0)| = 1$. From Theorem 85, there is a further sub-sequence of $k(l)$ such that $\frac{\mathbf{X}^{k'(l)}(|x_{k'(l)}|\tau)}{|x_{k'(l)}|} \to x(\tau)$ almost surely where the function $x(\cdot) \in S(x(0))$. Under the stability condition (A.12), we have that for any fluid-limit function $x(\cdot) \in S(x(0))$, $x(\tau) = 0$ whenever $|x(0)| \leq 1$. This establishes that given any arbitrary sequence of initial conditions x_k with $|x_k| \to \infty$, one can find a further sub-sequence $k'(l)$ such that

$$\lim_{k'(l)\to\infty} \frac{1}{|x_{k'(l)}|}|\mathbf{X}^{k'(l)}(|x_{k'(l)}|\tau)| = 0, \ a.s. \tag{A.21}$$

Therefore, we can conclude that for any sequence x_k with $|x_k| \to \infty$, we have $\frac{1}{|x_k|}|\mathbf{X}^k(|x_k|\tau)|$ tends to 0 in probability. But since, the family of random

208

variables $\left\{ \frac{|\mathbf{X}_x(|x|\tau)|}{|x|} \right\}_{x \in \mathbf{N}^{N_\epsilon}\setminus\{0\}}$ is uniformly integrable, we have that

$$\lim_{k\to\infty} \frac{1}{|x_k|}\mathbf{E}[|\mathbf{X}^k(|x_k|\tau)|] = 0. \tag{A.22}$$

As x_k was an arbitrary sequence, Equation (A.18) holds whenever condition (A.12) holds. □

From Equation (A.18), we have that for any $\epsilon > 0$, there is a large enough A_ϵ such that Equation (A.14) holds. Furthermore, for any finite A, the set $\{x \in \mathbf{N}^{N_\epsilon} : ||x||_\infty \leq A\}$ is finite. Hence, we have that $\mathbf{X}(t)$ is stable under the stability condition (A.12) which proves Theorem 3.

Generalization to arbitrary Link Distance T

To generalize the proof for arbitrary link distances T, we need to construct the appropriate discretization of the chain $\phi_t^{(\epsilon)}$. Once, we construct an appropriate discrete state space chain, then it is easy to see that the fluid version of this chain will satisfy inequality (A.17) and Lemma 86 as is. This will conclude that the case with arbitrary link distance T also yields the same stability result.

The discrete state space process in this case will naturally involve two vectors $\{X_i(t)\}_{i=1}^{n_\epsilon}$ and $\{Y_i(t)\}_{i=1}^{n_\epsilon}$, which represent the vector of transmitters and receivers in the discrete grid. However, in addition, we need a list of vectors $\{\mathcal{M}_i(t)\}_{i=1}^{n_\epsilon}$ where $\mathcal{M}_i$ is a n_ϵ dimensional vector whose jth coordinate

denotes how many transmitters in cell i have a corresponding receiver in cell j. The triple $\mathbf{X}_t := (X_i(t), Y_i(t), \mathcal{M}_i(t))_{i=1}^{n_\epsilon}$ then evolves in a Markovian fashion on a countable state-space. The evolution of $\mathbf{X}(t)$ is as follows. To each cell i, a new receiver is born at rate $\lambda\epsilon^2$. When, a receiver is born in cell i, we first pick an uniformly random location in the cell $A_i \subset \mathbf{S}$ and then centered around this point, we draw a ball of radius T and pick the location of the transmitter uniformly on the circumference to decide the cell in which the transmitters land. Thus, at the instant of birth, both a transmitter and receiver is born. Thus, conditioned on the event that a receiver is placed in cell i, there is a distribution on the set $\{1, 2, \cdots, n_\epsilon\}$ from which we sample the cell to place the corresponding transmitter in. To compute the interference seen at any receiver, we sum up the interference power from all transmitters in $\{Y_i\}_{i=1}^{n_\epsilon}$ including the intended signaling transmitter, which forms an upper bound on the interference. On the event of a death of a receiver in cell i, we also delete an uniformly random transmitter such that it has a receiver in cell i.

This process $\mathbf{X}(t)$ can be studied using fluid limits as above but with significantly more computations. The fluid equations for this case (which is the analog of Theorem 85) will be as follows.

$$\frac{d}{dt}x_i = \lambda\epsilon^2 - \frac{x_i l(T)}{\sum_{j=1}^{n_\epsilon} y_i l_\epsilon(a_i, a_j)}$$

whenever $x_i > 0$, else $\frac{d}{dt}x_i = 0$. Since, the number of transmitters and receivers

are the same at all instants of time, we get the following inequality immediately

$$\frac{d}{dt}y_i \le \lambda\epsilon^2 - \frac{y_i l(T)}{||y||_\infty \sum_{j=1}^{n^\epsilon} l_\epsilon(a_i, a_j)}$$

whenever $y_i > 0$. This is an inequality and not an equality due to the fact that the interference is measured by a transmitter process $||y||_\infty \mathbf{1}$ which coordinate wise dominates the original transmitters y. Thus, we can see that by employing the Lyapunov function $L(z)$ for $z = (x, y, \mathcal{M})$ as $L(z) := ||y||_\infty$, we will get exactly the same inequality as in Equation (A.17). Furthermore, it is easy to check that Lemma 86 holds as is with $|\mathbf{X}(t)| := ||y||_\infty$. This will then establish that Theorem 83 will hold as is for the chain constructed in this paragraph with generalized link distance T, which concludes the proof.

$\square$

Appendix C

A.3   Proof of Theorem 5

Proof. The proof idea is to apply Rate-Conservation equations similar to that of Theorem 1. For any receiver-transmitter pair $(x; y) \in \phi_t$, define $B_t(x) = \sum_{T \in \phi_t^{tx} \backslash \{y\}} f(||T - x||)$ and the cadlag process $\mathcal{B}_t = \sum_{x \in \phi_t^{Rx}} B_t(x)$.

Since we assume that the dynamics ϕ_t is ergodic, we write RCL for the stochastic process $\mathcal{B}_t$

$$\lambda |S| \mathbb{E} \left[2 \int_{x \in \mathbf{S}} B_0(x) \frac{dx}{|\mathbf{S}|} \right] =$$

$$\lambda_d \mathbb{E} \left[\frac{\mathbf{R}_0}{\mathbb{E}[\mathbf{R}_0]} \sum_{T_n \in \phi_0} \frac{R(T_n, \phi_0)}{\mathbf{R}_0} 2 B_0(T_n) \right] \quad \text{(A.23)}$$

The LHS follows from PASTA and the fact that a birth can happen anywhere in $\mathbf{S}$ uniformly and independently. The RHS follows from the Papangelou's theorem that the point process on $\mathbb{R}$ corresponding to the death epochs admits $\mathbf{R}_t = \frac{1}{L} \sum_{X_n \in \phi_0} R(X_n, \phi_0)$ as its Stochastic Intensity with respect to the filtration $\mathcal{F}_t = \sigma \left(\{ \phi_s : s \leq t \} \right)$, the sigma algebra generated by the location of the links. We also have $\lambda_d = \lambda |S|$ from Equation (A.1) and $\mathbb{E}[\mathbf{R}_0] = \lambda |S|$ from Equation (A.2). Using this to simplify Equation (A.23), we get

$$\mathbb{E}[B_0(0)] = \frac{1}{\lambda |S| L} \mathbb{E} \left[\sum_{T_n \in \phi_0} R(T_n, \phi_0) B_0(T_n) \right], \quad \text{(A.24)}$$

where we used Fubini's theorem and the fact that ϕ_0 is stationary in simplifying the LHS. Using the definition of Palm probability to simplify the RHS, we get

$$\mathbb{E}[B_0(0)] = \frac{\beta |\mathbf{S}|}{\lambda L |\mathbf{S}|} \mathbb{E}^0_{\phi_0} \left[R(0, \phi_0) B_0(0) \right]. \quad \text{(A.25)}$$

Since both $f(\cdot)$ and the path-loss $l(\cdot)$ are positive non-increasing functions, we have the deterministic behavior that if $B_0(0)$ increases, then $R(0, \phi_0)$ decreases. Hence, we can use the association inequality

$$\mathbb{E}^0_{\phi_0} \left[R(0, \phi_0) B_0(0) \right] \leq \mathbb{E}^0_{\phi_0} \left[R(0, \phi_0) \right] \mathbb{E}^0_{\phi_0} \left[B_0(0) \right] \quad \text{(A.26)}$$

Employing Inequality (A.26) in Equation (A.25), and the RCL $\lambda L = \beta \mathbb{E}^0_{\phi_0}[R(0, \phi_0)]$ from equation (A.3) we get

$$\mathbb{E}[B_0(0)] \leq \mathbb{E}^0_{\phi_0}[B_0(0)]. \tag{A.27}$$

$\square$

Appendix D

A.4 Proof of Theorem 85

Proof. This can be argued by contradiction. Assume that for some $\epsilon > 0$ and a sub-sequence

$$\mathbb{P}\left(\inf_{f \in S(x(0))} \sup_{t \in [0,T]} |z_k^{-1} X(z_k t) - f(t)| > \epsilon\right) \geq \epsilon \tag{A.28}$$

Without loss of generality, assume the above holds true for all $k \geq 1$.

The trajectories of the process $X^k(t)$ can be written in terms of independent unit-rate Poisson process A_i^k and D_i^k

$$X_i^k(t) = X_i^k(0) + A_i^k(\lambda \epsilon^2 t) - $$
$$D_i^k \left(\int_0^t X_i^k(u) \log_2\left(1 + \frac{1}{N_0 + I_i^\epsilon(t)} du\right)\right). \tag{A.29}$$

That is, $X^k(t)$ is a functional of the Point Process satisfying the set of Equations (A.29).

One can rewrite equation (A.29) by a change of variables as

$$\frac{1}{z_k} X_i^k(z_k t) = \frac{1}{z_k} X_i^k(0) + \frac{1}{z_k} A_i^k(\lambda \epsilon^2 z_k t) -$$
$$\frac{1}{z_k} D_i^k \left(\int_0^{z_k t} X_i^k(u) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(t)} \right) du \right). \quad (A.30)$$

Now replacing u by $z_k l$, we get the following

$$\frac{1}{z_k} X_i^k(z_k t) = \frac{1}{z_k} X_i^k(0) + \frac{1}{z_k} A_i^k(\lambda \epsilon^2 z_k t) -$$
$$\frac{1}{z_k} D_i^k \left(z_k \int_0^t X_i^k(z_k l) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(z_k l)} \right) dl \right), \quad (A.31)$$

which can be written as

$$\frac{1}{z_k} X_i^k(z_k t) = \frac{1}{z_k} X_i^k(0) + \lambda \epsilon^2 t -$$
$$\int_0^t X_i^k(z_k l) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(z_k l)} \right) dl + \delta_i^k(t), \quad (A.32)$$

where the error term $\delta_i^k(t)$ satisfies the stochastic bound

$$\sup_{t \in [0,T]} |\delta_i^k(t)| \leq \frac{1}{z_k} \sup_{t \in [0,\lambda \epsilon^2 T]} |A_i^k(z_k t) - z_k t| +$$
$$\frac{1}{z_k} \sup_{t \in [0,T \log_2(e)]} |D_i^k(z_k t) - z_k t|. \quad (A.33)$$

The error term $\delta_i^k(t)$ is bounded by the following lemma.

Lemma 87. [261] Let Ξ be a unit rate PPP on the real line. Then for all $T > 0$ and all $\lambda > 0$,

$$\mathbb{P} \left(\sup_{t \in [0,T]} |\Xi(t) - t| \geq \lambda T \right) \leq e^{-Th(\lambda)} + e^{-Th(-\lambda)} \quad (A.34)$$

where $h(\lambda) = (1 + \lambda) \log(1 + \lambda) - \lambda$.

This lemma in particular implies that there exists a sub-sequence $k(l), l \geq 1$ and a sequence $\epsilon(l) \to 0$ such that $\forall i$

$$\sum_{l \geq 1} \mathbb{P} \left(\sup_{t \in [0,T]} |\delta_i^{k(l)}(t)| \geq \epsilon(l) \right) < \infty$$

By Borel-Cantelli's lemma, there exists a sub sequence such that for all i, $\lim_{l \to \infty} \sup_{t \in [0,T]} |\delta_i^{k(l)}(t)| \to 0$ almost surely.

Now consider the random function

$w_k(t) = \int_0^t X_{ij}^k(z_k l) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(z_k l)} \right) dl$ which is Lipschitz for each sample path ω, i.e.

$$w_k(t) - w_k(s) = \int_s^t X_{ij}^k(z_k l) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(z_k l)} \right) dl \tag{A.35}$$

$$\leq (t - s) \frac{\log_2(e)}{\sup_{x,y \in \mathbf{S}} l^\epsilon(x, y)} < \infty. \tag{A.36}$$

From the Arzela-Ascoli theorem, there exists a sub-sequence such that $w_k(t)$ converges uniformly on $[0, T]$ to a Lipschitz continuous function $D_i(t)$ for each sample path ω. This along with the bound on $\sup_{t \in [0,T]} |\delta_{ij}(t)|$ yields that there is a sub-sequence such that

$$\frac{X_i^k(z_k t)}{z_k} \to x_{ij}(t) := x_i(0) + \lambda_i t - D_i(t), \ \ a.s., \tag{A.37}$$

where the convergence happens uniformly over $[0, T]$. $D_i(t)$ is Lipschitz since $x_i(t)$ is Lipschitz continuous. It remains to show that $\dot{D}_i(t) = \frac{x_i(t)}{I_i(t)}$. Since $D_i(t)$

215

is Lipschitz continuous, by Rademacher's theorem, it is differentiable almost everywhere on $[0, T]$. For all $h > 0$,

$$\int_t^{t+h} X_i^k(z_k l) \log_2 \left(1 + \frac{1}{N_0 + I_i^\epsilon(z_k l)} \right) dl \to$$

$$\int_t^{t+h} \frac{x_i(l)}{I_i^{\epsilon, f}(l)} dl.$$

This follows from dominated convergence and the Lipschitz continuity of $l \to x_{ij}(l)$. Therefore $\dot{D}_i(t) = \frac{x_i(t)}{I_i^{\epsilon, f}(t)}$.

Hence, we have shown that given any sequence of initial conditions $X^k(0)$ and number z_k such that the limit $\frac{X^k(0)}{z_k} = x(0)$ exists, we can find a sub-sequence k_l such that $\frac{X^{k_l}(z_{k_l} t)}{z_{k_l}}$ converges almost surely to the Lipschitz continuous fluid limit function $x(t)$. This is a contradiction and hence the theorem is proved. $\square$

Appendix B

Proofs from Chapter 3

In this chapter, we present the proof of Theorem 19. The proof is spread out in three sections and organized as follows. In Section B.1, we present the main stability theorem on the torus systems. This is similar to the proof of Theorem 3 in Chapter 2 and hence we only outline the sketch. Subsequently in Section B.2, we write down rate-conservation equations for the torus systems and establish the formula for the mean queue length. This section forms the core technical innovation in this thesis about this model. Among many things, we show a form of *tightness* of the marginal steady-state queue length with respect to the truncations of space. Subsequently in Section B.3, we use this tightness to conclude that the steady stte of the infinite space dynamics is in a precise sense the limit of the stationary distributions of steady state queue lengths in space truncated finite space systems.

B.1   Space Truncated Finite Systems

In this section, we discuss a finite version of the infinite queueing network. For any $n \in \mathbb{Z}_+$, we consider two n-truncated systems, both of which are obtained by restricting the dynamics to the set $B_n(0)$, the l_∞ ball of ra-

dius n centered at 0. For notational convenience, we shall drop 0 and denote by $B_n := B_n(0)$ the l_∞ ball of radius n centered at 0. For every $n \in \mathbb{N}$, we define two truncated dynamics, $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ and $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$. The process $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ evolves with the set B_n 'wrapped around' to form a torus. More precisely, the process $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ is driven by $(\mathcal{A}_i, \mathcal{D}_i)_{i \in B_n}$. The arrival dynamics is the same as before where each queue receives customers as a rate λ PPP independent of everything else. In other words, for all $i \in B_n$, at each epoch of $\mathcal{A}_i$, a customer is added to queue i. The departure dynamics is driven by $\mathcal{D}_i$ as before, but we treat the set B_n as a torus. More precisely, given any $i, j \in B_n$, define $d_n(i,j) := (i-j) \mod n$, where the modulo operation is coordinate-wise. Thus, at any time t, and any $i \in B_n$, the rate at which a departure occurs from queue i at time t in the process $\{y_i^{(n)}(t)\}_{i \in B_n(0)}$ is $\frac{y_i^{(n)}(t)}{\sum_{j \in B_n} a_{d_n(i,j)} y_j^{(n)}(t)}$. Since n is finite, the stochastic process $\mathbf{y}^{(n)}(t)$ is a continuous time Markov process on a countable state-space, i.e., on $\mathbb{N}^{(2n+1)^d}$. Moreover, since the jumps are triggered by a finite number of Poisson processes, this chain has almost surely no-explosions.

Similarly, the process $\{z_i^{(n)}(t)\}_{i \in B_n}$ is driven by the arrival data $(\mathcal{A}_i, \mathcal{D}_i)_{i \in B_n}$ as before, but this time the set B_n is viewed as a subset of $\mathbb{Z}^d$ and in particular the 'edge effects' are retained. The arrival rate to any queue $i \in B_n$ in the system $\{z_i^{(n)}(t)\}_{i \in B_n}$ is λ, while there are no arrivals to queues in B_n^{c}, i.e., an arrival rate of 0. Moreover, the queue lengths of queues in B_n^{c} is set to 0, i.e., for all $t \geq 0$ and all $i \in B_n^{\mathsf{c}}$, we have $z_i^{(n)}(t) = 0$. The departure process for

any queue $i \in B_n$, is identical to the original infinite system. At any time $t \geq 0$, and any $i \in B_n$, the rate of departure from queue i at time t is given by $\frac{z_i^{(n)}(t)}{\sum_{j \in \mathbb{Z}^d} a_{i-j} z_j^{(n)}(t)}$. From the monotonicity in the dynamics, we have the following proposition.

Proposition 88. For all $n > L$, there exists a coupling of the processes $\{x_i(\cdot)\}_{i \in \mathbb{Z}^d}$, $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$ and $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ such that for all $t \in \mathbb{R}$, and all $i \in \mathbb{Z}^d$, we have $x_i(t) \geq z_i^{(n)}(t)$ and $y_i^{(n)}(t) \geq z_i^{(n)}(t)$ almost surely.

The following property of the truncated systems will be used in the analysis of the infinite system.

Theorem 89. For all $n > L$ and $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, the Markov process $\{y_i^{(n)}(t)\}_{i \in B_n(0)}$ is positive recurrent. Let $\pi^{(n)}$ denote the stationary queue length distribution on $\mathbb{N}$ of any queue $i \in B_n(0)$ and let Z be distributed as $\pi^{(n)}$. Then there exists a $c > 0$ possibly depending on n such that $\mathbb{E}[e^{cZ}] < \infty$.

Remark 90. The symmetry in the torus implies that the marginal stationary queue length distribution of any queue i, $\pi^{(n)}$, is the same for all i.

Remark 91. The existence of an exponential moment yields that all power moments of $\pi^{(n)}$ are finite.

Remark 92. In view of Proposition 88, if $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then for all $n \in \mathbb{N}$, the process $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$ is positive recurrent. Moreover, for all $i \in B_n$, the stationary distribution of $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$, denoted by $\{\tilde{\pi}_i^{(n)}\}_{i \in B_n}$, is such that there exists a $c > 0$ possibly depending on n satisfying $\mathbb{E}[e^{cZ_i}] < \infty$, where Z_i is distributed according to $\tilde{\pi}_i^{(n)}$.

219

Proof Sketch

We provide a sketch here and defer the details to Section 5 of the extended version in [338] The proof is technical with a lot of details based on monotonicity and the standard properties of a single-server queue with light-tailed service time distributions. So we present its summary just to highlight the key ideas involved. To prove the theorem, we will define a modified dynamics $\{\tilde{y}_i^{(n)}(t)\}_{t \geq 0, i \in B_n(0)}$ which is coupled with the evolution of $\{y_i^{(n)}(t)\}_{i \in B_n(0)}$. We construct the modified dynamics such that it satisfies $\tilde{y}_i^{(n)}(t) \geq y_i^{(n)}(t)$ a.s. for all $i \in B_n(0)$ and $t \geq 0$. We do this by discretization of continuous time to discrete by choosing sufficiently small interval h, i.e. times $\cdots, -h, 0, h, 2h, \cdots$ will form time slot boundaries. We then restrict departures so that at-most one departure can occur in a time period. We also modify the arrivals so that in any time slot, the difference between the maximum number of arrivals and the minimum number of arrivals in a time slot is at-most a constant. From monotonicity, the dynamics with such modified arrivals and departures can be coupled to provide an upper bound to the true queue lengths. We describe in detail this construction in Section 5 of [338]. We further identify a large r, and *equalize* the queues after every r time-slots, i.e. at times $\cdots, -rh, 0, rh, 2rh, \cdots$, we add fictitious customers so that all queues have the same number of customers. If the number of customers is smaller than a constant y_0, we further add more customers till every queue has at least y_0 customers. Thus, at the end of every r time-slots, every queue has the same number of customers which is at least y_0. From a coupling argument, we show

that after the addition of the fictitious customers, the queue length follows the trajectory of an appropriately modified $GI/GI/1$ queue which is stable. Thus, we have dominated our process $\{y_i^{(n)}(t)\}_{i \in B_n}$ so that every one of them is dominated from above by a stable $GI/GI/1$ queue with light-tailed service time distributions, and hence the stationary distribution of $\{y_i^{(n)}(t)\}_{i \in B_n}$ is also light-tailed.

B.2 Rate Conservation Arguments

This section forms the central tool used in our analysis. We shall consider the space truncated systems introduced before to explicitly write down differential equations for certain functionals of the dynamics. The key result we will establish in this section is a closed form formula for the mean of the steady state queue length of the space truncated torus system. To do so, we will use the general rate conservation principle of Palm calculus [49] to derive certain relations between the system parameters in steady state. We shall assume $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$ throughout in this section. We shall let $n > L$ be arbitrary and fixed for the rest of this section. In this section, we shall again consider the two space truncated stochastic processes $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ and $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$ to be in steady-state. Recall that the process $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ is one wherein the set B_n is viewed as a torus with its end points identified and the process $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$ is one with the end effects, i.e., with the set B_n is viewed as a subset of $\mathbb{Z}^d$. Furthermore, we denote by $\pi^{(n)}$, the steady state distribution of $y_0^{(n)}(0)$ and by the translation invariance on the torus, the steady state distribution of $y_i^{(n)}(0)$,

for all $i \in B_n$. Similarly, for all $i \in B_n$, we shall denote by $\tilde{\pi}_i^{(n)}$, the steady state distribution of the marginal $z_i^{(n)}(t)$. Notice that the marginal distributions in the process $z_i^{(n)}(\cdot)_{i \in B_n}$ depend on the coordinate, unlike in the process $y_i^{(n)}(\cdot)_{i \in B_n}$. For notational simplicity, we shall denote by $\mu^{(n)}$, the mean of $\pi^{(n)}$ and for all $i \in B_n$, by $\nu_i^{(n)}$, the mean of $\tilde{\pi}_i^{(n)}$. In this section, we shall study two stochastic processes - $\{\mathbb{I}_t\}_{t \in \mathbb{R}}$ and $\{\tilde{\mathbb{I}}_t\}_{t \in \mathbb{R}}$, with $\mathbb{I}_t := y_0^{(n)}(t)(\sum_{j \in \mathbb{Z}^d} a_j y_j^{(n)}(t))$ and $\tilde{\mathbb{I}}(t) := \sum_{i \in B_n} z_i^{(n)}(t)(\sum_{j \in \mathbb{Z}^d} a_j z_{i-j}^{(n)}(t))$. If one were to be more precise, one should use the notation $\mathbb{I}_t^{(n)}$ and $\tilde{\mathbb{I}}_t^{(n)}$, but we drop the superscript n to simplify the notation. In words, the process $(\{\mathbb{I}_t\})_{t \in \mathbb{R}}$ considers the interference seen at a typical queue in $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$, the system where the set B_n is viewed as a torus and $(\{\tilde{\mathbb{I}}_t\})_{t \in \mathbb{R}}$ considers the total interference in the process $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$, the system with boundary effects, where the set B_n is viewed as a subset of $\mathbb{Z}^d$. Observe that since B_n is a torus, the marginals of the process $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ are identical and hence, we can consider a typical queue, but as the marginals of $\{z_i^{(n)}(\cdot)\}_{i \in B_n}$ are different due to the edge effects, we study the total inter- ference at all queues instead of the interference seen at a typical queue. Since $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, and the systems $\{y_i^{(n)}(t)\}_{i \in B_n}$ and $\{z_i^{(n)}(t)\}_{i \in B_n}$ are in steady state, and the queue lengths in both systems posses exponential moments, it is the case that for all $t \in \mathbb{R}$, $\mathbb{E}[\mathbb{I}_t] < \infty$ and $\mathbb{E}[\tilde{\mathbb{I}}(t)] < \infty$.

The main technical results of this section are Propositions 93, 94 and Lemma 100. These will then help us to derive closed form expressions for the mean queue length and a bound on the second moment for the original infinite

system.

Proposition 93.

$$\frac{d}{dt}\mathbb{E}[\mathbb{I}_t] = 0 = \lambda a_0 + 2\lambda\left(\sum_{j\in B_n} a_j\right)\mu^{(n)} -$$

$$\mathbb{E}\left[R(0)\left(a_0(2y_0^{(n)}(0) - 1) + \sum_{i\in B_n\backslash\{0\}} a_i y_i^{(n)}(0)\right) + \sum_{i\in B_n\backslash\{0\}} R(i)a_i y_0^{(n)}(0)\right],$$

$$\text{(B.1)}$$

where for any $i \in B_n$,

$$R(i) := \frac{y_i^{(n)}(0)}{\sum_{j\in B_n} a_{d_n(i,j)} y_j^{(n)}(0)}.$$

Proof. We provide a heuristic derivation of the differential equation using the PASTA property of the arrival and departure process and skip all the technical details as it is standard. For example see the Appendix of [56] for an account of the derivation. In a small interval of time Δt, in every queue, there will be exactly one arrival with probability roughly $\lambda\Delta t$. The chance that two or more arrivals occur in a time interval Δt in the entire network is $\mathrm{O}\big((\Delta t)^2\big)$, where the $\mathrm{O}(\cdot)$ hides all system parameters (for ex. λ, n) other than Δt as they are fixed. On an arrival at queue 0, the increase in the quantity $\mathbb{I}_0$ is $\mathbb{E}[(y_0^{(n)}+1)(a_0(y_0^{(n)}+1)+\sum_{j\in B_n\backslash\{0\}} a_j y_j^{(n)})-y_0^{(n)}(\sum_{j\in B_n} a_j y_j^{(n)})]$, which is equal to $\mathbb{E}[a_0 + \sum_{j\in B_n} a_j y_j^{(n)}]$. Similarly, the average increase in $\mathbb{I}_0$ due to an arrival in the neighboring queues of 0 is $\mathbb{E}[(y_0^{(n)})(a_i(y_i^{(n)} + 1) + \sum_{j\in B_n\backslash\{i\}} a_j y_j^{(n)}) - y_0^{(n)}(\sum_{j\in B_n} a_j y_j^{(n)})]$, which is equal to $\mathbb{E}[a_i y_0^{(n)}]$. The chance that there are two

223

or more arrivals is $\mathrm{O}\big((\Delta t)^2\big)$, which is small. Thus, the average increase due to arrivals is $\lambda \Delta t \mathbb{E}[a_0(y_0^{(n)} + 1) + \sum_{j \in B_n} a_j y_j^{(n)} + \sum_{j \in B_n \backslash \{0\}} a_j y_0^{(n)}] + \mathrm{O}\big((\Delta t)^2\big)$. After simplification, and using the fact that the variables $y_j^{(n)}$ all have the same mean, we get that the average increase in time Δt is

$$\lambda \Delta t \left(a_0 + 2\mu^{(n)} \left(\sum_{j \in B_n} a_j \right) \right) + \mathrm{O}\big((\Delta t)^2\big). \tag{B.2}$$

Likewise, with probability $R(i)\Delta t$, there will be a departure from queue i. When a customer leaves from queue 0, which occurs with probability $R(0)\Delta t$ the average decrease in $\mathbb{I}_0$ is then $\mathbb{E}[(a_0((y_0^{(n)})^2 - a_0(y_0^{(n)} - 1)^2 + \sum_{i \in B_n \backslash \{0\}} a_i y_i^{(n)})]$. Similarly, a departure from queue i, which occurs with probability $R(i)\Delta t$ results in an average decrease in $\mathbb{I}_0$ of $\mathbb{E}[a_i y_0^{(n)}]$. The chance that two or more possible departures occur in time Δt is $\mathrm{O}\big((\Delta t^2)\big)$, which is small. Thus, the total average decrease in $\mathbb{I}_0$ due to departures is

$$\Delta t \mathbb{E}\left[R(0) \left(a_0(2y_0^{(n)} - 1) + \sum_{i \in B_n \backslash \{0\}} a_i y_i^{(n)} \right) + \sum_{i \in B_n \backslash \{0\}} R(i) a_i y_0^{(n)} \right] + \mathrm{O}\big((\Delta t)^2\big). \tag{B.3}$$

Hence, we see from Equations (B.2) and (B.3), that

$$\frac{1}{\Delta t} \mathbb{E}[\mathbb{I}(t + \Delta t) - \mathbb{I}(t)] = \lambda \left(a_0 + 2\mu^{(n)} \left(\sum_{j \in B_n} a_j \right) \right) - $$

$$\mathbb{E}\left[R(0) \left(a_0(2y_0^{(n)} - 1) + \sum_{i \in B_n \backslash \{0\}} a_i y_i^{(n)} \right) + \sum_{i \in B_n \backslash \{0\}} R(i) a_i y_0^{(n)} \right] + \mathrm{O}\big(\Delta t\big).$$

The proposition is concluded by letting Δt go to 0.

$\square$

Now, we compute the differential equation for the space truncated system, by carefully taking into consideration the 'edge effects' introduced by the truncation to the set B_n. Denote by the set $B_n^{(I)} \subset B_n$, where $B_n^{(I)} := \{z \in B_n : \forall y \text{ s.t. } ||y - z||_\infty \leq L, y \in B_n\}$. In words, the set $B_n^{(I)}$ is the set of all points $z \in B_n$ such that the l_∞ ball of radius L is completely contained in B_n.

Proposition 94.

$$\frac{d}{dt}\mathbb{E}[\tilde{\mathbb{I}}(t)] = 0 \geq -2(1 - \lambda \sum_{j\in\mathbb{Z}^d} a_j) \sum_{i\in B_n^{(I)}} \nu_i^{(n)} + 2\lambda a_0 |B_n| - 2 \sum_{i\in B_n\setminus B_n^{(I)}} \nu_i^{(n)}.$$

Proof. A rigorous proof of this is standard and we skip it. For example see [56]. Instead we outline the computations required in establishing this proposition. Furthermore to lighten the notation in the proof, we drop the superscript n, as n is fixed and does not change in the course of the proof. Thus, we shall denote $\{z_i^{(n)}(t)\}_{i\in B_n}$ by $\{z_i(t)\}_{i\in B_n}$ and the steady-state means by $(\nu_i)_{i\in B_n}$, instead of $(\nu_i^{(n)})_{i\in B_n}$. As in the proof of Proposition 93, we shall consider a small interval Δt of time such that at most one event of either an arrival or departure occurs anywhere in the network in the set B_n. Roughly speaking, with probability $\lambda\Delta t$, there will be an arrival in some queue $i \in B_n$. In the rest of the proof, we shall partition the set B_n into $B_n^{(I)}$ and $B_n \setminus B_n^{(I)}$.

From similar computations as in the previous proposition, if there is an arrival in queue $i \in B_n^{(I)}$, the increase in $\tilde{\mathbb{I}}(t)$ will be

$$\mathbb{E}\bigg[(z_i(t){+}1)(a_0(z_i(t){+}1){+}\sum_{j\in\mathbb{Z}^d\backslash\{0\}} a_j z_{i-j}(t)){-}z_i(t)(z_i(t){+}\sum_{j\in\mathbb{Z}^d\backslash\{0\}} a_j z_{i-j}(t)){+}$$

$$\sum_{l\in B_n\backslash\{i\}}(z_l(t)(a_{i-l}(z_i(t){+}1){+}\sum_{j\in\mathbb{Z}^d\backslash\{i\}} a_{j-l}z_j(t)){-}z_l(t)(a_{i-l}z_i(t)\sum_{j\in\mathbb{Z}^d\backslash\{i\}} a_{j-l}z_j(t)))\bigg].$$

This follows since if there is an extra customer in queue i, then the total interference is increased both at queue i and any other queue j such that $a_{i-j} > 0$. From the PASTA property, we know that at the moment of arrival, $\{z_i(t)\}_{I\in B_n}$ is in steady-state and in particular, $\mathbb{E}[z_i(t)] = \nu_i$. Thus, the above expression can be simplified as

$$\sum_{j\in\mathbb{Z}^d} a_j\nu_{i-j} + a_0 + a_0\nu_i + \sum_{l\in B_n\backslash\{i\}}\nu_l a_{i-l}.$$

If $i \in B_n^{(I)}$, the above expression is equal to

$$2\sum_{j\in\mathbb{Z}^d} a_j\nu_{i-j} + a_0,$$

while if $i \in B_n^{(I)}\setminus B_n$, we use the trivial inequality

$$\sum_{j\in\mathbb{Z}^d} a_j\nu_{i-j} + a_0 + a_0\nu_i + \sum_{l\in B_n\backslash\{i\}}\nu_l a_{i-l} \geq a_0.$$

Thus, the average increase in the time interval Δt in the interference $\mathbb{I}(t)$ due to an arrival event is at least

$$\lambda\Delta t\left(\sum_{i\in B_n^{(I)}}(2\sum_{j\in\mathbb{Z}^d} a_j\nu_{i-j} + a_0) + \sum_{i\in B_n\backslash B_n^{(I)}} a_0\right) + \mathrm{O}(\Delta t^2).$$

226

Since for all $i \in B_n^{(I)}$, the l_∞ ball of radius L is contained within the set B_n, we can further simplify the above expression as

$$\lambda \Delta t \left(\sum_{i \in B_n^{(I)}} 2\nu_i \sum_{j \in \mathbb{Z}^d} a_j + a_0 \right) + \lambda a_0 \Delta t |B_n \setminus B_n^{(I)}| + \mathrm{O}(\Delta t^2).$$

Similarly, we can compute the average decrease in $\tilde{\mathbb{I}}(t)$ due to a departure event. Roughly, the probability of a departure from any queue $i \in B_n$ is given by $R_i(t)$ where $R_i(t) = \frac{z_i(t)}{\sum_{j \in \mathbb{Z}^d} a_{i-j} z_j(t)}$. If there is a departure from queue i, the average decrease can be computed as

$$\mathbb{E}\Bigg[(z_i(t))(a_0 z_i(t) + \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j z_{i-j}(t)) - (z_i(t) - 1)(a_0(z_i(t) - 1) + \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j z_{i-j}(t))$$

$$+ \sum_{l \in B_n \setminus \{i\}} (z_l(t)(a_{i-l} z_i(t) + \sum_{j \in \mathbb{Z}^d \setminus \{i\}} a_{j-l} z_j(t)) - z_l(t)(a_{i-l}(z_i(t) - 1) \sum_{j \in \mathbb{Z}^d \setminus \{i\}} a_{j-l} z_j(t)))\Bigg].$$

We do not need to worry about the fact that $z_i(t) - 1$ can be negative since, in this case, the rate of departure $R_i(t)$ will be 0. Thus the average rate of decrease in the interference due to a departure can be written as

$$\Delta t \left(\sum_{i \in B_n} \mathbb{E}[R_i(t)(a_0 z_i(t) + \sum_{j \in \mathbb{Z}^d} a_j z_{i-j}(t) - a_0) + \sum_{l \in B_n \setminus \{i\}} a_{i-l} z_l(t)] \right) + \mathrm{O}(\Delta t^2).$$

Using the fact that for all $i \in B_n$ and all $t \in \mathbb{R}$, we have $R_i(t)(\sum_{j \in \mathbb{Z}^d} a_j z_{i-j}(t)) = z_i(t)$, we can simplify the average rate of decrease as

$$\Delta t (\sum_{i \in B_n} 2\nu_i - a_0 \sum_{i \in B_n} \mathbb{E}[R_i(t)]) + \mathrm{O}(\Delta t^2).$$

227

However, as $\{z_i(\cdot)\}_{i \in B_n}$ is a stationary process, $\sum_{i \in B_n} \mathbb{E}[R_i(t)] = \lambda |B_n|$. This then gives that average rate of change in $\mathbb{E}[\mathbb{I}(t)]$ is

$$\frac{1}{\Delta t} \mathbb{E}[\mathbb{I}(t\Delta t) - \mathbb{I}(t)] \geq$$
$$\lambda\left(\sum_{i \in B_n^{(I)}} 2\nu_i\left(\sum_{j \in \mathbb{Z}^d} a_j\right) + a_0\right) + \lambda a_0 |B_n \setminus B_n^{(I)}| - \sum_{i \in B_n} 2\nu_i + a_0 \sum_{i \in B_n} \lambda + \mathrm{O}(\Delta t).$$

Letting Δt go to 0, we obtain the bound in Proposition 94. $\qquad \square$

We now state Lemma 95 which holds as a consequence of the rate conservation argument. This establishes a closed form expression for the mean queue length in steady state in the space truncated torus system $\{y_i^{(n)}(t)\}_{i \in B_n}$. Recall that the system $\{y_i^{(n)}(t)\}_{i \in B_n}$ is in steady state. Thus, the stochastic process $(\mathbb{I}_t)_{t \in \mathbb{R}}$ is stationary. In particular, $\frac{d}{dt}\mathbb{E}[\mathbb{I}_t]$ is equal to 0. Thus, from Proposition 93, we have the following key lemma

Lemma 95. For all $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$ and all $n > L$,

$$\mu^{(n)} = \frac{\lambda a_0}{1 - \left(\sum_{j \in \mathbb{Z}^d} a_j\right)\lambda}. \tag{B.4}$$

Remark 96. Note that we assumed $a_0 = 1$ in the model. For completeness, we give the derivation for any general $a_0 > 0$.

This lemma in particular yields that the mean number of customers in the steady state of the space truncated torus is independent of n, provided n is large enough. This in particular gives $\sup_n \mu^{(n)} < \infty$.

Proof. From Equation (B.1), we get

$$\lambda \left(a_0 + 2\mu^{(n)} \left(\sum_{j \in \mathbb{Z}^d} a_j \right) \right)$$

$$= \mathbb{E}\left[R(0) \left((a_0(2x_0^{(n)} - 1) + \sum_{i \in B_n \setminus \{0\}} a_i x_i^{(n)} \right) + \sum_{i \in B_n \setminus \{0\}} R(i) a_i x_0^{(n)} \right].$$

Now, we use the following version of the Mass Transport Principle for unimodular random graphs (see also [255]):

Proposition 97. The following formula holds.

$$\mathbb{E}\left[\sum_{i \in B_n \setminus \{0\}} R(i) a_i y_0^{(n)} \right] = \mathbb{E}\left[\sum_{i \in B_n \setminus \{0\}} R(0) a_i y_i^{(n)} \right].$$

Proof. The proof follows from the standard argument of Mass Transport involving swapping double sums. Observe from the definition of the dynamics, the queue lengths $\{y_k^{(n)}\}_{k \in B_n}$ is translation invariant on the torus B_n. Hence, for all $j \in B_n$, the variables $y_j^{(n)} \sum_{i \in B_n \setminus \{j\}} R(i) a_{i-j}$ are identically distributed, and in particular have the same means. The proposition is now proved thanks

to the following calculations.

$$\mathbb{E}\left[\sum_{i\in B_n\setminus\{0\}} R(i)a_i y_0^{(n)}\right] = \frac{1}{|B_n|}\mathbb{E}\left[\sum_{j\in B_n} y_j^{(n)} \sum_{i\in B_n\setminus\{j\}} R(i)a_{i-j}\right]$$

$$\overset{(a)}{=} \frac{1}{|B_n|}\mathbb{E}\left[\sum_{i\in B_n} R(i) \sum_{j\in B_n\setminus\{i\}} a_{i-j}y_j^{(n)}\right]$$

$$\overset{(b)}{=} \frac{1}{|B_n|}\mathbb{E}\left[\sum_{i\in B_n} R(i) \sum_{j\in B_n\setminus\{i\}} a_{j-i}y_j^{(n)}\right]$$

$$\overset{(c)}{=} \mathbb{E}\left[\sum_{i\in B_n\setminus\{0\}} R(0)a_i y_i^{(n)}\right].$$

Equality (a) follows by swapping the order of summations, which is licit since they each contain finitely many terms. Equality (b) follows since $a_k = a_{-k}$ for all $k \in \mathbb{Z}^d$. Equality (c) again follows from the fact that for all $i \in B_n$, $R(i)\sum_{j\in B_n\setminus\{i\}} a_{j-i}y_j^{(n)}$ are identically distributed. This is a consequence of the queue lengths $\{y_k^{(n)}\}_{k\in B_n}$ being translation invariant on the torus. $\qquad\square$

We now show how to conclude the proof of Lemma 95, using the conclusions of Proposition 97. Intuitively, Proposition 97 can be interpreted by considering the finite graph with vertices on the torus B_n with a directed edge from i to j in B_n with weight $R(i)a_{d_n(i-j)}y_j^{(n)}$. This random graph, when rooted in 0, is unimodular and hence the Mass Transport Principle holds ([255]). Since $a_i = a_{-i}$, we get that the average decrease is $\mathbb{E}[-a_0 R(0) + 2R(0)\sum_{i\in B_n} a_i y_i^{(n)}]$. Now, $\mathbb{E}[R(0)] = \lambda$, and since, for all $i \in B_n$, $\mathbb{E}[y_i^{(n)}] = \mu^{(n)}$,

$$2\lambda(\sum_{i\in B_n} a_i)\mu^{(n)} + 2\lambda a_0 = \mathbb{E}[2R(0)(\sum_{i\in B_n} a_i y_i^{(n)})]. \tag{B.5}$$

230

But since $R(0)(\sum_{i \in B_n} a_i y_i^{(n)}) = y_0^{(n)}$, we get

$$\mu^{(n)} = \frac{\lambda a_0}{1 - (\sum_{i \in \mathbb{Z}^d} a_i)\lambda}.$$

<div style="text-align: right">□</div>

Corollary 98. If $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then the sequence of probability measures $\{\pi^{(n)}\}_{n>L}$ is tight.

Proof. From Markov's inequality, we have

$$\mathbb{P}[X > Q] \leq \frac{\lambda a_0}{Q(1 - (\sum_{i \in \mathbb{Z}^d} a_i)\lambda)},$$

where X is distributed according to $\pi^{(n)}$. Thus, for every $\varepsilon > 0$, we can find Q large such that $\sup_{n>L} \mathbb{P}_{\pi^{(n)}}[X > Q] < \varepsilon$. □

B.2.1 Finiteness of Second Moments

In this section, we establish that under the conditions stated in Proposition 21, the second moments of the marginals of the queue lengths of $\{y_i^{(n)}(\cdot)\}_{i \in B_n}$ are uniformly bounded in n. In order to show this, we need the following auxiliary lemma. For completeness, we provide expressions without assuming that the value of a_0 of the interference sequence $\{a_i\}_{i \in \mathbb{Z}^d}$ to be 1.

Lemma 99. For all $\lambda > 0$, $\{a_i\}_{i \in \mathbb{Z}^d}$, $d \in \mathbb{N}$ and $n > L$, we have $\mathbb{E}[y_0^2 \sum_{i \in B_n} R_i a_i] \leq 2c\mathbb{E}[y_0^2]$, where the constant c equals $\frac{\sqrt{a_0^2 + a_0 \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j} - a_0}{\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j}$.

Proof. From symmetry, i.e., $a_i = a_{-i}$ for all $i \in B_n$ and translation invariance on the torus, we have

$$\mathbb{E}[y_0^2 \sum_{i \in B_n} R_i a_i] = \mathbb{E}[R_0 \sum_{i \in B_n} y_i^2 a_i]. \tag{B.6}$$

Let $c > 0$ be such that $2ca_0 = a_0 - (\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_i)c^2$. The only positive solution to this equation is $c = \frac{\sqrt{a_0^2 + a_0 \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j} - a_0}{\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j}$. Thus, we have the following chain of equations -

$$\sum_{i \in B_n} a_i y_i^2 = a_0 y_0^2 + \sum_{i \in \mathbb{Z}^d \setminus \{0\}} a_i y_i^2 = 2ca_0 y_0^2 + \sum_{i \in \mathbb{Z}^d \setminus \{0\}} a_i(y_i^2 + c^2 y_0^2) \geq 2cy_0 \sum_{i \in B_n} a_i y_i, \tag{B.7}$$

where the last inequality follows from $y_i^2 + c^2 y_0^2 \geq 2cy_0 y_i$. Thus, from Equations (B.6) and (B.7), we have

$$\mathbb{E}[R_0 \sum_{i \in B_n} y_i^2 a_i] \geq 2c\mathbb{E}[R_0 y_0 \sum_{i \in B_n} a_i y_i] = 2c\mathbb{E}[y_0^2].$$

$\square$

Lemma 100. For all $\lambda < \frac{2}{3}\frac{1+c}{\sum_{j \in \mathbb{Z}^d} a_j}$, we have $\mathbb{E}[(y_0^{(n)})^2] \leq \frac{2\mu(\lambda + \lambda \sum_{j \in \mathbb{Z}^d} a_j + 1)}{2(1+c) - 3\lambda \sum_{j \in \mathbb{Z}^d} a_j}$, where the constant c is $\frac{\sqrt{a_0^2 + a_0 \sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j} - a_0}{\sum_{j \in \mathbb{Z}^d \setminus \{0\}} a_j}$.

Proof. The proof of this lemma is an application of the rate conservation equation to the process $(y_0^{(n)})^2 I_0^{(n)}$, where $I_0^{(n)}$ is the interference given by $I_0^{(n)} = \sum_{j \in \mathbb{Z}^d} a_j y_j^{(n)}$. For brevity of notation, we remove the superscript n in the calculations. The average increase in the process $y_0^2 I_0$ due to an arrival is

232

given by

$$\mathbb{E}\left[\lambda\left(((y_0+1)^2(I_0+1)-y_0^2 I_0)+\sum_{j\in\mathbb{Z}^d\setminus\{0\}}(y_0^2(I_0+a_j)-y_0^2 I_0)\right)\right]$$

$$=\mathbb{E}\left[\lambda\left(((y_0^2+2y_0+1)(I_0+1)-y_0^2 I_0)+\sum_{j\in\mathbb{Z}^d\setminus\{0\}}y_0^2 a_j\right)\right]$$

$$=\mathbb{E}\left[\lambda\left(y_0^2+2y_0 I_0+2y_0+I_0+1+\sum_{j\in\mathbb{Z}^d\setminus\{0\}}y_0^2 a_j\right)\right]$$

$$=\lambda\sum_{j\in\mathbb{Z}^d}a_j\mathbb{E}[y_0^2]+2\lambda\mathbb{E}[y_0 I_0]+2\lambda\mu+\lambda\mu\sum_{j\in\mathbb{Z}^d}a_j+\lambda$$

$$\leq 3\lambda\sum_{j\in\mathbb{Z}^d}a_j\mathbb{E}[y_0^2]+2\lambda\mu+\lambda\mu\sum_{j\in\mathbb{Z}^d}a_j+\lambda. \tag{B.8}$$

In the last simplification, we use the bound that $y_0 y_j \leq \frac{1}{2}(y_0^2+y_j^2)$ and the fact that $\mathbb{E}[y_0^2]=\mathbb{E}[y_j^2]$ for all $j\in B_n$. Similarly, the average decrease in the process $(y_0^{(n)})^2 I_0^{(n)}$ due to a departure is then given by

$$\mathbb{E}\left[R_0\left(y_0^2 I_0-(y_0-1)^2(I_0-1)\right)+\sum_{j\in\mathbb{Z}^d\setminus\{0\}}R_j(y_0^2 I_0-y_0^2(I_0-a_j))\right]$$

$$=\mathbb{E}\left[R_0\left(y_0^2 I_0-(y_0^2-2y_0+1)(I_0-1)\right)+\sum_{j\in\mathbb{Z}^d\setminus\{0\}}R_j y_0^2 a_j\right]$$

$$=\mathbb{E}\left[R_0\left(y_0^2+2y_0 I_0-2y_0-I_0+1\right)\sum_{j\in\mathbb{Z}^d\setminus\{0\}}R_j y_0^2 a_j\right]$$

$$=\mathbb{E}\left[y_0^2\sum_{j\in\mathbb{Z}^d}a_j R_j\right]+2\mathbb{E}[R_0 y_0 I_0]-2\mathbb{E}[R_0 y_0]-\mathbb{E}[R_0 I_0]+\mathbb{E}[R_0]. \tag{B.9}$$

Since the process $\{y_i^{(n)}\}_{i\in B_n}$ is stationary, the average change due to arrivals and departures is 0, i.e., the difference between the left hand sides of Equations

(B.8) and (B.9) equals 0 . Further, using the simplifications that $\mathbb{E}[R_0] = \lambda$, $R_0 I_0 = y_0$ and $R_0 \leq 1$ almost surely, we have by taking a difference of Equations (B.8) and (B.9), that

$$0 \leq 3\lambda \sum_{j \in \mathbb{Z}^d} a_j \mathbb{E}[y_0^2] + 2\lambda\mu + \lambda\mu \sum_{j \in \mathbb{Z}^d} a_j + \lambda -$$
$$\left(\mathbb{E}\left[y_0^2 \sum_{j \in \mathbb{Z}^d} a_j R_j \right] + 2\mathbb{E}[R_0 y_0 I_0] - 2\mathbb{E}[R_0 y_0] - \mathbb{E}[R_0 I_0] + \mathbb{E}[R_0] \right).$$

The above equation can be simplified by employing the result of Lemma 99 as follows:

$$0 \overset{(a)}{\leq} 3\lambda \sum_{j \in \mathbb{Z}^d} a_j \mathbb{E}[y_0^2] + 2\lambda\mu + \lambda\mu \sum_{j \in \mathbb{Z}^d} a_j - 2c\mathbb{E}[y_0^2] - 2\mathbb{E}[y_0^2] + 2\mu + \lambda\mu \sum_{j \in \mathbb{Z}^d} a_j,$$
$$\leq 2\mu(\lambda + \lambda \sum_{j \in \mathbb{Z}^d} a_j + 1) - (2(1 + c) - 3\lambda \sum_{j \in \mathbb{Z}^d} a_j)\mathbb{E}[y_0^2].$$

The inequality (a) follows from Lemma 99. By rewriting the last display, it is clear that if $\lambda < \frac{2(1+c)}{3} \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then $\mathbb{E}[y_0^2] \leq \frac{2\mu(\lambda + \lambda \sum_{j \in \mathbb{Z}^d} a_j + 1)}{2(1+c) - 3\lambda \sum_{j \in \mathbb{Z}^d} a_j}$.

$\square$

The above proposition in particular gives us the following corollary.

Corollary 101. For all $n > L$, if $\lambda < \frac{2(1+c)}{3} \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then $\sup_{n \geq L} \mathbb{E}[(y_0^{(n)})^2] < \infty$.

Based on discrete event simulations, we conjectured in the initial version of this paper posted online, that the second moment is uniformly bounded

in n for the entire stability region. This was subsequently established by [356].
See also Conjecture 30 in Section 3.3.1, and the discussions following it.

B.3 Coupling From the Past - Proofs of Theorem 19 and Proposition 21

The key idea is to use monotonicity and the backward coupling representation. In order to implement the proof, we need some additional notation. For any $T > 0$ and $n \in \mathbb{N}$ such that $n > L$, and any $i \in \mathbb{Z}^d$, we define the random variables $x_{i;T}(0), y_{i;T}^{(n)}(0)$ and $z_{i;T}^{(n)}(0)$. These variables represent the number of customers in queue i at time 0 in three different dynamics which will be coupled and driven by the same arrival and departure processes - $(\mathcal{A}_i, \mathcal{D}_i)_{i \in \mathbb{Z}^d}$. In all of them, the subscript i refers to queue i and T refers to the fact that the system started empty at time $-T$. We now describe the three different dynamics in question:

1. $x_{i;T}(0)$ denotes the number of customers in queue i at time 0 in the original infinite dynamics.

2. $y_{i;T}^{(n)}(0)$ denotes the number of customers in queue i at time 0 for the dynamics restricted to the set $B_n(0)$ *viewed as a torus*. Hence $\{y_{i;T}^{(n)}(0)\}_{i \in B_n(0)}$ is the queue length of the process studied in Section B.1.

3. $z_{i;T}^{(n)}(0)$ denotes the number of customers at time 0 for the dynamics restricted set B_n, not seen as a torus. Thus for all $i \in B_n(0)^c$, we have $z_{i;T}^{(n)}(0) = 0$, by definition.

The following two propositions follow immediately from monotonicity.

Proposition 102. For all $T > 0$, all $n > L$, and all $i \in \mathbb{Z}^d$, we have $x_{i;T}(0) \geq z_{i;T}^{(n)}(0)$ and $y_{i;T}^{(n)}(0) \geq z_{i;T}^{(n)}(0)$ almost surely.

Proposition 103. For all $n > L$, almost surely, the following limits exist:

$$x_{i;\infty}(0) := \lim_{T \to \infty} x_{i;T}(0),$$
$$y_{i;\infty}^{(n)}(0) := \lim_{T \to \infty} y_{i;T}^{(n)}(0),$$
$$z_{i;\infty}^{(n)}(0) := \lim_{T \to \infty} z_{i;T}^{(n)}(0).$$

Note that the distribution of the random variable $y_{0;\infty}^{(n)}$ is the marginal on coordinate 0 of the probability measure $\boldsymbol{\pi}^{(n)}$, whose existence was proved in Theorem 89 We also established in Corollary 98 that the sequence of probability measures $\{\pi^{(n)}\}_{n \in \mathbb{N}}$ is tight. Moreover, in view of Lemma 16, it suffices to show that queue 0 is stable to conclude that the entire network is stable. Hence for notational brevity, we will omit the queue and time index by adopting the following simplified notation for the rest of this section: $x_T := x_{0;T}(0)$, $y_T^{(n)} := y_{i;T}^{(n)}(0)$, $z_T^{(n)} := z_{i;T}^{(n)}(0)$, where $T \in [0, \infty]$.

Proposition 104. Almost surely, for every $T \geq 0$, we have $\lim_{n \to \infty} z_T^{(n)} = x_T$.

Proof. Notice from the construction (outlined in [338]), that for every finite T, there exists a random subset $X \subset \mathbb{Z}^d$ which is almost surely finite and such

that the value of x_T can be obtained by restricting the dynamics to the set X in the time interval $[-T, 0]$. Let N be any integer such that X is contained in B_n. Then, for all $n \geq N$, $x_T = z_T^{(n)}$. □

Lemma 105. The sequence $z_\infty^{(n)}$ is non-decreasing in n and almost surely converges to a finite integer valued random variable denoted by $z_\infty^{(\infty)}$.

Proof. Note that for all finite T, $z_T^{(n)}$ is non-decreasing in n. Thus for any $n > m$, we have $z_T^{(n)} \geq z_T^{(m)}$, for all T. Now, taking a limit in T on both sides, which we know exist from Proposition 103, we see that $z_\infty^{(n)} \geq z_\infty^{(m)}$. This establishes the fact that $z_\infty^{(n)}$ is an non-decreasing sequence and hence the almost sure limit $\lim_{n \to \infty} z_\infty^{(n)} := z_\infty^{(\infty)}$ exists. We now show the finiteness of $z_\infty^{(\infty)}$. Note that for all n and T, $z_T^{(n)} \leq y_T^{(n)}$. Now, taking a limit in T, we see that $z_\infty^{(n)} \leq y_\infty^{(n)}$. The distribution of the random variable $y_\infty^{(n)}$ is the probability measure $\pi^{(n)}$ on $\mathbb{N}$. From Corollary 98, the sequence $\{\pi_n\}$ is tight. Let $\tilde{\pi}^{(n)}$, $n \in \mathbb{N}$, denote the distribution of $z^{(n)}$. Thus the sequence $\{\tilde{\pi}^{(n)}\}_{n \in \mathbb{N}}$ is tight as well since $z_\infty^{(n)} \leq y_\infty^{(n)}$ almost surely. Moreover, due to monotonicity, $z_\infty^{(n)}$ converges almost surely to a random variable $z_\infty^{(\infty)}$. But since the sequence $\{\tilde{\pi}^{(n)}\}_{n \in \mathbb{N}}$ is tight, the limiting random variable $z_\infty^{(\infty)}$ is almost surely finite. □

Lemma 106. There exists a random $N \in \mathbb{N}$, such that for all $n \geq N$, there exists a random $T_n \in \mathbb{R}^+$, such that for all $t \geq T_n$, $z_\infty^{(\infty)} = z_t^{(n)}$.

Proof. From the previous lemma, $z_\infty^{(n)}$ converges almost surely to a finite limit as $n \to \infty$. Since the random variables $\{z_\infty^{(n)}\}_{n \in \mathbb{N}}$ are integer valued, there

exists a random N such that $z_\infty^{(\infty)} = z_\infty^{(n)}$, $\forall n \geq N$. Now, since, for each T and n, $z_T^{(n)}$ is integer valued, the existence of an almost surely finite limit $\lim_{T \to \infty} z_T^{(n)}$ implies that there exists a T_n, almost surely finite and such that $z_t^{(n)} = z_\infty^{(n)}$ for all $t \geq T_n$. Now, combining the two, for every $n \geq N$, we can find a T_n such that $z_t^{(n)} = z_\infty^{(n)}$ for all $t \geq T_n$. Since N is such that for all $n \geq N$, $z_\infty^{(n)} = z_\infty^{(\infty)}$, the lemma is proved. $\qquad\square$

Lemma 107. Let T_N be the random variable defined in Lemma 106. For all $t \geq T_N$, we have $x_t = z_\infty^{(\infty)}$.

Proof. Let $m \geq N$ and $t \geq T_N$ be arbitrary. Observe that $\lim_{T \to \infty} z_T^{(m)} = z_\infty^{(m)} = z_\infty^{(\infty)}$, where the second equality follows from the fact that $m \geq N$. From Lemma 106, there exists an almost surely finite T_m such that for all $t \geq T_m$, we have $z_t^{(m)} = z_\infty^{(m)} = z_\infty^{(\infty)}$. Let $t' \geq \max(t, T_m)$. Since $t' \geq T_m$, we have $z_{t'}^{(m)} = z_\infty^{(\infty)}$. Basic monotonicity gives us the following two inequalities:

$$z_t^{(m)} \geq z_t^{(n)} = z_\infty^{(\infty)},$$
$$z_t^{(m)} \leq z_{t'}^{(m)} = z_\infty^{(\infty)}.$$

The first inequality follows from monotonicity in space and the second from monotonicity in time. Thus, $z_t^{(m)} = z_\infty^{(\infty)}$. But since $m \geq N$ was arbitrary, it must be the case that $x_t = \lim_{m \to \infty} z_t^{(m)} = z_\infty^{(\infty)}$, where the first equality follows from Proposition 104. Thus we have established that, for all $t \geq T_N$, we have $x_t = z_\infty^{(\infty)}$ and, in particular, $x_\infty = \lim_{t \to \infty} x_t = z_\infty^{(\infty)}$ is an almost surely finite random variable.

$\qquad\square$

Corollary 108. If $\lambda < \frac{1}{\sum_{i \in \mathbb{Z}^d} a_i}$, then the following interchange of limits holds true:

$$\lim_{t \to \infty} \lim_{n \to \infty} z_t^{(n)} = \lim_{n \to \infty} \lim_{t \to \infty} z_t^{(n)} = x_\infty = z_\infty^{(\infty)} < \infty \text{ a.s.}$$

Corollary 109. If $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then $\mathbb{E}[x_\infty] \leq \frac{\lambda a_0}{1 - \lambda(\sum_{j \in \mathbb{Z}^d} a_j)} < \infty$.

Proof. From Corollary 108, $x_\infty = \lim_{n \to \infty} z_\infty^{(n)}$. Moreover since $z_\infty^{(n)}$ is non-decreasing in n, it follows from the monotone convergence theorem that $\mathbb{E}[x_\infty] = \lim_{n \to \infty} \mathbb{E}[z_\infty^{(n)}]$. As $z_\infty^{(n)} \leq y_\infty^{(n)}$ and $\sup_{n \geq L} \mathbb{E}[y_\infty^{(n)}] = \frac{\lambda a_0}{1 - \lambda(\sum_{j \in \mathbb{Z}^d} a_j)}$ from Lemma 95, we get $\mathbb{E}[x_\infty] \leq \frac{\lambda a_0}{1 - \lambda(\sum_{j \in \mathbb{Z}^d} a_j)} < \infty$. $\qquad\square$

Now to finish the proof of Theorem 19, we need to conclude about the mean queue length value, which we do in the following Lemma.

Lemma 110. If $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$, then $\mathbb{E}[x_{0;\infty}(0)] \geq \frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j}$.

Proof. We shall choose $n > L$ arbitrary and consider the stochastic process $\tilde{\mathbb{I}}^{(n)}(t)$ defined in Proposition 94. We shall let $\tilde{\mathbb{I}}^{(n)}(t)$ be stationary as $\lambda < \frac{1}{\sum_{j \in \mathbb{Z}^d} a_j}$. Furthermore, notice from Theorem 89 that the truncated process $\{z_i^{(n)}(t)\}_{i \in B_n}$ has exponential moments. Thus, we have for all $n \in \mathbb{N}$ and all $t \in \mathbb{R}$, $\mathbb{E}[\tilde{\mathbb{I}}^{(n)}(t)] < \infty$. Thus, we can equate $\frac{d}{dt}\mathbb{E}[\tilde{\mathbb{I}}^{(n)}(t)]$ to 0 in Proposition 94 along with the fact $|B_n| \geq |B_n^{(I)}|$, to obtain

$$0 \geq -2(1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j) \sum_{i \in B_n^{(I)}} \nu_i^{(n)} + 2\lambda |B_n^{(I)}| - 2 \sum_{i \in B_n \setminus B_n^{(I)}} \nu_i^{(n)}. \qquad \text{(B.10)}$$

Re-arranging the inequality, we see that

$$\frac{1}{|B_n^{(I)}|} \sum_{i \in B_n^{(I)}} \nu_i^{(n)} \geq \frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j} - \frac{\sum_{i \in B_n \setminus B_n^{(I)}} \nu_i^{(n)}}{|B_n^{(I)}|}. \qquad \text{(B.11)}$$

Notice that $\nu_i^{(n)} \leq \mathbb{E}[x_{0;\infty}(0)]$ which in turn thanks to Corollary 108 is upper bounded by $\frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j}$. Further-more, from elementary counting arguments we have $\lim_{n \to \infty} |B_n \setminus B_n^{(I)}| |B_n^{(I)}| = 0$. Thus, we obtain for all $n > L$,

$$\mathbb{E}[x_{0;\infty}(0)] \geq \frac{1}{|B_n^{(I)}|} \sum_{i \in B_n^{(I)}} \nu_i^{(n)} \geq \frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j} \left(1 - \frac{|B_n \setminus B_n^{(I)}|}{|B_n^{(I)}|} \right). \qquad \text{(B.12)}$$

Taking a limit as $n \to \infty$ concludes the proof.

$\square$

B.3.1 Proof of Proposition 21

From Corollaries 101, and 108, the conclusion of Proposition 21 follows.

B.4 Proof of Proposition 20 - Uniqueness of Stationary Solution

To carry out the proof, we shall employ the following rate conservation principle.

Lemma 111. If $\{q_i\}_{i \in \mathbb{Z}^d}$ is a stationary solution to the dynamics satisfying $\mathbb{E}[q_0^2] < \infty$, then $\mathbb{E}[q_0] = \frac{\lambda a_0}{1 - \lambda \sum_{j \in \mathbb{Z}^d} a_j}$.

Proof. Since $\mathbb{E}[q_0^2] < \infty$, then we can apply the same proof verbatim of Proposition 93, where the stochastic process $\mathbb{I}(t) := q_0(t) \sum_{i \in \mathbb{Z}^d} a_i q_i(t)$. Then the conclusion of Proposition 93 and Lemma 95 follow.

$\square$

We now prove Proposition 20 with the aid of certain monotonicity arguments.

Proof. Let $\boldsymbol{\pi}'$ be a stationary measure on $(\mathbb{Z}^d)^{\mathbb{N}}$, with finite second moment for the marginals. Let this distribution be different from $\boldsymbol{\pi}$, the distribution corresponding to the minimal stationary solution $\{x_{i;\infty}(0)\}_{i \in \mathbb{Z}^d}$. We show by elementary coupling and monotonicity arguments that $\boldsymbol{\pi} = \boldsymbol{\pi}'$. Let $T > 0$ be arbitrary. We couple the evolutions of the two systems $\{y_{i;T}(\cdot)\}_{i \in \mathbb{Z}^d}$ and $\{x_{i;T}(\cdot)\}_{i \in \mathbb{Z}^d}$ as follows: Let $\{q_i\}_{i \in \mathbb{Z}^d}$ be distributed according to $\boldsymbol{\pi}'$, independently of everything else. Let $\{y_{i;T}(-T)\}_{i \in \mathbb{Z}^d}$ be such that $y_{i;T}(-T) = q_i$, for all $i \in \mathbb{Z}^d$ and $\{x_{i;T}(-T)\}_{i \in \mathbb{Z}^d}$ be empty, i.e., for all $i \in \mathbb{Z}^d$, we have $x_{i;T}(-T) = 0$. Thus, for all $i \in \mathbb{Z}^d$, $x_{i;T}(-T) \leq y_{i;T}(-T)$. Monotonicity in Lemma 14 implies that, almost surely, for all $i \in \mathbb{Z}^d$, we have $x_{i;T}(0) \leq y_{i;T}(0)$. By the definition of invariance, $\{y_{i;T}(0)\}_{i \in \mathbb{Z}^d}$ is distributed as $\boldsymbol{\pi}'$ with $\mathbb{E}[y_{0;T}(0)]$ given in Lemma 111. From Proposition 107, we know that as $T \to \infty$, $x_{0;T}(0)$ converges almost surely to a random variable which has a finite first moment. Furthermore, from the hypothesis of the proposition, we know that the almost sure limit $\lim_{T \to \infty} x_{0;T}(0)$ also possesses finite second moment. Thus from the dominated convergence theorem, we have that $\lim_{T \to \infty} \mathbb{E}[x_{0;T}(0)] = \mathbb{E}[x_{0;\infty}(0)]$,

which is also the same as given in Lemma 111. Thus $\boldsymbol{\pi}'$ coordinate-wise dominates $\boldsymbol{\pi}$. But they have the same first moment. This implies that the two probability measures are the same. $\qquad\square$

Appendix C

Proofs from Chapter 4

C.0.1 Proof of Theorem 37

Proof. We need the following lemma first.

Lemma C.0.1. *Let Y be any $\mathcal{E}$-valued R.V. and $g(\cdot) : C \times \mathcal{E} \to \mathbb{R}$.*

$$\mathbb{E}[\sup_{x \in C} g(x, Y)] \geq \sup_{x \in C} \mathbb{E}[g(x, Y)].$$

Proof. We have

$$\sup_{x \in C} g(x, Y) \geq g(x, Y) \ \forall x \in C,$$

and hence

$$\mathbb{E}[\sup_{x \in C} g(x, Y)] \geq \mathbb{E}[g(x, Y)] \ \forall x \in C. \tag{C.1}$$

Since (C.1) is valid for all $x \in C$, we can pick the supremum on the RHS, i.e.

$$\mathbb{E}[\sup_{x \in C} g(x, Y)] \geq \sup_{x \in C} \mathbb{E}[g(x, Y)]. \tag{C.2}$$

$\square$

We can now prove Theorem 37.

$$\mathcal{R}_{I_2}^{\pi^*} = \mathbb{E}[\max_{i \in [1,T]} \sup_{j \geq 1} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_{I_2}]]$$

$$\stackrel{(a)}{=} \mathbb{E}[\mathbb{E}[\max_{i \in [1,T]} \sup_{j \geq 1} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_{I_2}]|\mathcal{F}_{I_1}]]$$

$$\stackrel{(b)}{\geq} \mathbb{E}[\max_{i \in [1,T]} \sup_{j \geq 1} \mathbb{E}[\mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_{I_2}]|\mathcal{F}_{I_1}]]$$

$$\stackrel{(c)}{=} \mathbb{E}[\max_{i \in [1,T]} \sup_{j \geq 1} \mathbb{E}[p_i(\mathrm{SINR}_0^{i,j})|\mathcal{F}_{I_1}]] = \mathcal{R}_{I_1}^{\pi^*}$$

where (a) follows from the tower property of expectation, (b) follows from Lemma C.0.1 and (c) follows from the tower property of expectation and the fact that $\mathcal{F}_{I_1} \subseteq \mathcal{F}_{I_2}$. $\square$

C.0.2   Proof of Lemma 4.3.1

Proof. From the properties of PPP, we know that almost-surely, the distances $\{r_k^i\}_{k \geq 1}$ are distinct i.e. satisfy $r_k^i > r_{k+1}^i$. Denoting $S_k = P_i l_i(r_k^i)$ for each $k \in \mathbb{N}$ (instead of representing it as S_k^i, we drop the i in this proof for simplicity), we can write (4.3) as

$$j_i = \arg\sup_{j \geq 1} \mathbb{E}\left[p_i \left(\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d} \right) \bigg| \mathcal{F}_I \right]$$

$$\stackrel{(a)}{=} \arg\sup_{j \geq 1} \mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{i \neq j} S_i H_i} \bigg| \mathcal{F}_I \right],$$

$$\tag{C.3}$$

where (a) follows from the fact that the function $p_i(\cdot)$ is non-decreasing.

$$\sup_{j \geq 1} \mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\mathcal{F}_I\right]$$

$$= \sup_{j \geq 1} \mathbb{E}\left[\mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\sigma(\mathcal{F}_I \cup \phi_i)\right]\bigg|\mathcal{F}_I\right]$$

$$\leq \mathbb{E}\left[\sup_{j \geq 1} \mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\sigma(\mathcal{F}_I \cup \phi_i)\right]\bigg|\mathcal{F}_I\right], \qquad \text{(C.4)}$$

where the inequality follows from Lemma C.0.1. Since conditioned on ϕ_i, we have S_k deterministic and H_k conditionally i.i.d. given ϕ_i and independent of $\mathcal{F}_I$, we have

$$\sup_{j \geq 1} \mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\sigma(\mathcal{F}_I \cup \phi_i)\right]$$

$$= \sup_{j \geq 1} \mathbb{E}\left[S_j H_j|\sigma(\mathcal{F}_I \cup \phi_i)\right]\mathbb{E}\left[\frac{1}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\sigma(\mathcal{F}_I \cup \phi_i)\right]$$

$$= \sup_{j \geq 1} \mathbb{E}\left[S_j H_j|\phi_i\right]\mathbb{E}\left[\frac{1}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\phi_i\right]. \qquad \text{(C.5)}$$

Thus $j = 1$ achieves the supremum in (C.5) since $S_k > S_{k+1}$ and is deterministic given ϕ_i. Combining this fact with (C.4), we have

$$\sup_{j \geq 1} \mathbb{E}\left[\frac{S_j H_j}{N_0^i + \sum_{d \neq j} S_d H_d}\bigg|\mathcal{F}_I\right] \leq \mathbb{E}\left[\frac{S_1 H_1}{N_0^i + \sum_{d \geq 2} S_d H_d}\bigg|\mathcal{F}_I\right], \qquad \text{(C.6)}$$

which yields that $j_i = 1$. $\qquad \square$

C.0.3 Proof of Theorem 38

Proof. For ease of notation, we denote the path-loss function as simply $l(\cdot)$ instead of $l^{(\alpha)}(\cdot)$, i.e. implicitly assume the dependence on α as $l(x) = x^{-\alpha}$.

For each fixed α, we have from (4.2)

$$
\begin{aligned}
\pi_\alpha^*(0) \\
= \arg\max_{i\in[1,T]} \mathbb{E}\left[p\left(\frac{P_i l(r_1^i)H_1^i}{\sum_{j\geq 2} P_i l(r_j^i)H_j^i} \right) \middle| (r_1^l, \cdots, r_k^l)_{l=1}^T \right] \\
= \arg\max_{i\in[1,T]} \mathbb{E}\left[\frac{P_i l(r_1^i)H_1^i}{\sum_{j\geq 2} P_i l(r_j^i)H_j^i} \middle| (r_1^l, \cdots, r_k^l)_{l=1}^T \right] \\
= \arg\max_{i\in[1,T]} \mathbb{E}\left[\frac{1}{\sum_{j\geq 2} \frac{l(r_j^i)H_j^i}{l(r_1^i)H_1^i}} \middle| (r_1^i, \cdots, r_k^i) \right].
\end{aligned}
\tag{C.7}
$$

We now argue that for each technology i, the conditional expectation in (C.7) can be written as $A\frac{l(r_1^i)}{l(r_2^i)} - e_i^{(\alpha)}$ such that $e_i^{(\alpha)} \xrightarrow{\alpha\to\infty} 0$ almost surely and A is a positive constant independent of i and α. If we show this, then the lemma can be proved as follows:

$$
\pi_\alpha^*(0) = \arg\max_{i\in[1,T]} A\frac{l(r_1^i)}{l(r_2^i)} - e_i^{(\alpha)}
\tag{C.8}
$$

$$
\overset{(a)}{=} \arg\max_{i\in[1,T]} A\frac{r_2^i}{r_1^i} - e_i^{(\alpha)}
$$

$$
\xrightarrow[a.s.]{\alpha\to\infty} \arg\max_{i\in[1,T]} A\frac{r_2^i}{r_1^i} = \arg\max_{i\in[1,T]} \frac{r_2^i}{r_1^i},
\tag{C.9}
$$

where step (a) follows from the fact that $\frac{l(a)}{l(b)} = l(a/b)$ and the fact that $l(\cdot)$ is non-increasing. Since $e_i^{(\alpha)}$ converges to 0 almost-surely $\forall i \in [1, T]$, a finite set, we have uniform convergence i.e. $\sup_{i\in[1,T]} e_i^{(\alpha)} \xrightarrow{\alpha\to\infty} 0$ almost-surely which gives (C.9). In the rest of the proof, we show that (C.7) can be written as $A\frac{l(r_1^i)}{l(r_2^i)} - e_i^{(\alpha)}$.

Expanding on the conditional expectation in (C.7) using simple algebra to factor out the leading term, we get

$$\mathbb{E}\left[\frac{1}{\sum_{j\geq 2}\frac{l(r_j^i)H_j^i}{l(r_1^i)H_1^i}}\Big|(r_1^i,..,r_k^i)\right] \overset{(b)}{=} \frac{l(r_1^i)}{l(r_2^i)}\mathbb{E}\left[\frac{H_1^i}{H_2^i}\Big|(r_1^i,..,r_k^i)\right] -$$
$$\frac{l(r_1^i)}{l(r_2^i)}\mathbb{E}\left[\frac{H_1^i}{H_2^i}\left(\frac{Q_i^{(\alpha)}}{1+Q_i^{(\alpha)}}\right)\Big|(r_1^i,\cdots,r_k^i)\right], \quad \text{(C.10)}$$

where $Q_i^{(\alpha)} = \sum_{j\geq 3}\frac{l(r_j^i)H_j^i}{l(r_2^i)H_2^i}$. Step (b) follows from the fact that H_j^i are i.i.d. random-variables. Indeed (C.10) resembles (C.8) with the constant $A = \mathbb{E}\left[\frac{H_2^i}{H_1^i}\right]$ (which is independent of i and α). It thus remains to prove that the second term (which is the error $e_i^{(\alpha)}$) in (C.10) goes to 0 almost surely as α goes to infinity.

From Campbell's Theorem, we know that

$$\mathbb{E}[Q_i^{(\alpha)}|(r_1^i,\cdots,r_k^i),H_2^i] = \frac{\mathbb{E}[H]\sum_{z=3}^{k}l(r_z^i)}{H_2^i l(r_2^i)} +$$
$$\frac{\mathbb{E}[H]}{H_2^i l(r_2^i)}2\pi\lambda_i\int_{u\geq r_k^i}l(u)udu. \quad \text{(C.11)}$$

with the notation that $\sum_{z=a}^{b}\cdot = 0$ if $b < a$. Furthermore, since $l(x) = x^{-\alpha}$, we have $\frac{1}{l(\epsilon)}\int_{u\geq\epsilon}l(u)udu$ goes to 0 as α goes to ∞ for every $\epsilon > 0$. Thus, we have from (C.11) and the fact that for a homogeneous PPP of positive intensity λ_i, $r_j^i > r_{j+1}^i$ a.s. $\forall j \in \mathbb{N}$, we get

$$\lim_{\alpha\to\infty}\mathbb{E}[Q_i^{(\alpha)}|(r_1^i,\cdots,r_k^i),H_2^i] = 0 \text{ a.s.} \quad \text{(C.12)}$$

Note that we needed to invoke Campbell's theorem, since we need to conclude about a sum of infinite random variables involved in the definition of $Q^{(\alpha)}$. Thus,

$$
\begin{aligned}
e_i^{(\alpha)} &= \frac{l(r_1^i)}{l(r_2^i)} \mathbb{E}\left[\frac{H_1^i}{H_2^i} \left(\frac{Q_i^{(\alpha)}}{1 + Q_i^{(\alpha)}} \right) \middle| (r_1^i, \cdots, r_k^i) \right] \\
&\leq \frac{l(r_1^i)}{l(r_2^i)} \mathbb{E}\left[\frac{H_1^i}{H_2^i} \mathbb{E}[Q_i^{(\alpha)} | (r_1^i, \cdots, r_k^i), H_1^i, H_2^i] \middle| (r_1^i, \cdots, r_k^i) \right] \\
&\overset{(c)}{\underset{\alpha \to \infty}{\longrightarrow}} 0 \text{ a.s.}
\end{aligned} \tag{C.13}
$$

where step (c) follows from (C.12) (through Dominated Convergence) and the fact that H_1^i is a finite mean random variable independent of everything else. Since $e_i^{(\alpha)}$ is positive, inequality (C.13) yields that $e_i^{(\alpha)} \to 0$ a.s. $\qquad \square$

C.0.4 Proof of Lemma 4.5.1

Proof. From the definition of $f_i^*(\mathbf{r})$, we have,

$$
\begin{aligned}
f_i^*(\mathbf{r})\mathbf{dr} &= \mathbb{P}[r \in \mathbf{dr} | i = i^*] \\
&= \frac{\mathbb{P}[\{r \in \mathbf{dr}\} \cap \{i = i^*\}]}{\mathbb{P}[i = i^*]} \\
&\overset{(a)}{=} \frac{\mathbb{P}[\{r \in \mathbf{dr}\} \cap_{j \neq i} \{\pi_j(\mathbf{r}_j, \lambda_j) \leq \pi_i(\mathbf{r}, \lambda_j)\}]}{\mathbb{P}[i = i^*]} \\
&\overset{(b)}{=} f_i(\mathbf{r}) \prod_{j=1, j \neq i}^{T} F_{\pi_j}(\pi_i(\mathbf{r}, \lambda_i)) \frac{1}{p_i} \mathbf{dr},
\end{aligned} \tag{C.14}
$$

where p_i is the probability that $i = i^*$ and $\mathbf{dr}$ is an infinitesimal element of $\mathbb{R}^L$. Here (a) follows from the definition of i^* in (4.6) and (b) follows from the independence of the different point process and as a consequence independence of the observation vectors $\mathbf{r}_j$. $\qquad \square$

248

C.0.5 Proof of Theorem 39

Proof. The performance of a policy $\pi_i(\cdot)$ in (4.4) becomes:

$$
\begin{aligned}
\mathcal{R}_I^\pi &= \mathbb{E}[p_{i^*}(\text{SINR}_0^{i^*,j_{i^*}})] \\
&= \sum_{i=1}^T \mathbb{P}[i = i^*]\mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|i = i^*] \\
&= \sum_{i=1}^T \mathbb{P}[i = i^*]\mathbb{E}_{\mathbf{r}_i}[\mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}_i, i = i^*]] \\
&\overset{(a)}{=} \sum_{i=1}^T \mathbb{P}[i = i^*]\int_{\mathbf{r}\in\mathbb{R}^L} \mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}, i = i^*]f_i^*(\mathbf{r})dr \\
&\overset{(b)}{=} \sum_{i=1}^T \mathbb{P}[i = i^*]\int_{\mathbf{r}\in\mathbb{R}^L} \mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}]f_i^*(\mathbf{r})dr \\
&\overset{(c)}{=} \sum_{i=1}^T \int_{\mathbf{r}\in\mathbb{R}^L} \mathbb{E}[p_i(\text{SINR}_0^{i,j_i})|\mathbf{r}]f_i(\mathbf{r}) \prod_{j=1,j\neq i}^T F_{\pi_j}(\pi_i(\mathbf{r},\lambda_i))dr.
\end{aligned}
$$

We use the definition of $f_i^*(\mathbf{r})$ to perform the averaging over $\mathbf{r}_i$ on the event $i = i^*$ in step (a). Step (b) follows from the independence of ϕ_i across i and hence we can drop the conditioning on $i = i^*$. Step (c) follows from Lemma 4.5.1. □

C.0.6 Proof of Theorem 40

Proof. Consider first the case with information r_1^i:

$$c_p(1; r, \lambda, P, N_0, \beta)$$

$$= \mathbf{P}\left[\frac{Ph_1 r^{-\alpha}}{N_0 + \sum_{j\geq 2} Ph_j r_j^\alpha} > \beta \Big| r_1 = r\right]$$

$$\overset{(a)}{=} \exp\left(-\mu\beta N_0 P^{-1} r^\alpha\right) \mathbb{E}\left[\exp\left(-\mu\beta r^\alpha \sum_{j\geq 2} r_j^{-\alpha} h_j\right)\right],$$

$$\text{(C.15)}$$

where (a) follows follow from the fact that $\{h_i\}_{\{i\geq 1\}}$ are i.i.d. exponential random variables with mean $\frac{1}{\mu}$. Simplifying $\mathbb{E}\left[\exp\left(-\mu\beta r^\alpha \sum_{j\geq 2} r_j^{-\alpha} h_j\right)\right]$, we get

$$\mathbb{E}\left[\exp\left(-\mu\beta r^\alpha \sum_{j\geq 2} r_j^{-\alpha} h_j\right)\right] \overset{(b)}{=} \qquad\qquad \text{(C.16)}$$

$$\exp\left(-2\pi\lambda \int_{u\geq r} \left(1 - \mathbb{E}_h[e^{-h\mu\beta(\frac{u}{r})^{-\alpha}}]\right) u\,du\right)$$

$$\overset{(c)}{=} \exp\left(-2\pi\lambda \int_{u\geq r} \left(1 - \frac{\mu}{\mu\beta(\frac{u}{r})^{-\alpha} + \mu}\right) u\,du\right)$$

$$= \exp\left(-2\pi\lambda \int_{u\geq r} \left(\frac{1}{1 + \beta^{-1}(\frac{u}{r})^\alpha}\right) u\,du\right),$$

where (c) follow from the fact that $\{h_i\}_{\{i\geq 1\}}$ are i.i.d. exponential random variables with mean $\frac{1}{\mu}$. (b) follows from the expression for the Probability Generating Functional of an independently marked PPP and the fact that conditioned on the distance of the nearest point to the origin of a PPP of intensity λ as r_1 , the point process on $\mathbb{R}^2 \setminus B(0, r_1)$ is a homogeneous PPP with intensity λ.

The second case with information $[r_1^i, r_2^i]$ can be proven similarly:

$$c_p(1; [r_1, r_2], \lambda, P, N_0, \beta)$$

$$= \mathbf{P} \left[\frac{P h_1 r_1^{-\alpha}}{N_0 + P h_2 r_2^{-\alpha} + \sum_{j \geq 3} P h_j r_j^{\alpha}} > \beta \Big| r_1, r_2 \right]$$

$$\stackrel{(a)}{=} \exp\left(-\mu \beta N_0 P^{-1} r_1^{\alpha}\right) \mathbb{E}\left[\exp\left(-\mu \beta h_2 \left(\frac{r_1}{r_2}\right)^{\alpha}\right) \Big| r_1, r_2\right]$$

$$\mathbb{E}\left[\exp\left(-\mu \beta r_1^{\alpha} \sum_{j \geq 3} r_j^{-\alpha} h_j\right)\right].$$

$$\text{(C.17)}$$

The computation for $\mathbb{E}\left[\exp\left(-\mu \beta r_1^{\alpha} \sum_{j \geq 3} r_j^{-\alpha} h_j\right)\right]$ follows the steps similar to the above case and we skip it for brevity. We can compute $\mathbb{E}\left[\exp\left(-\mu \beta h_2 \left(\frac{r_1}{r_2}\right)^{\alpha}\right) \Big| r_1, r_2\right]$ since H_2 is an independent exponential random variable and hence that expectation is equal to $\left(\frac{1}{1 + \beta \left(\frac{r_1}{r_2}\right)^{\alpha}}\right)$. $\qquad \square$

C.0.7 Proof of Lemma 4.6.1

Proof.

$$F_{\pi_i}(y) = \mathbb{P}[c_p(1; r, \lambda_i, P_i, N_0, \beta) \leq y]$$

$$= \mathbb{P}\left[\int_{u=r}^{\infty} \frac{1}{1 + \beta^{-1} \left(\frac{u}{r}\right)^{\alpha}} u \, du \geq \frac{1}{2\pi \lambda_i} \ln\left(\frac{1}{y}\right)\right],$$

where the probability is with respect to the random variable r which is Rayleigh distributed with parameter $\frac{1}{\sqrt{2\pi \lambda_i}}$. In the above expression, making a change

of variables of $v = \frac{u}{r}$, we have

$$
\begin{aligned}
F_{\pi_i}(y) &= \mathbb{P}\left[r^2 \int_{v=1}^{\infty} \frac{1}{1 + \beta^{-1}(v)^{\alpha}} v dv \geq \frac{1}{2\pi\lambda_i} \ln\left(\frac{1}{y}\right)\right] \\
&= \mathbb{P}\left[r \geq \sqrt{\ln\left(\frac{1}{y}\right) \frac{1}{2\pi\lambda_i} \frac{1}{\int_{v=1}^{\infty} \frac{1}{1+\beta^{-1}(v)^{\alpha}} v dv}}\right] \\
&= e^{-\ln\left(\frac{1}{y}\right) \frac{1}{2}\left(\int_{v=1}^{\infty} \frac{1}{1+\beta^{-1}(v)^{\alpha}} v dv\right)^{-1}}.
\end{aligned}
\tag{C.18}
$$

$\square$

C.0.8 Proof of Lemma 4.6.2

Proof.

$$
\begin{aligned}
\mathbb{P}\left[\frac{r_2^i}{r_1^i} \leq x\right] &= \mathbb{E}[\mathbb{E}[\mathbf{1}(r_2^i \leq xr_1^i)|r_1^i]] \\
&\overset{(a)}{=} \mathbb{E}[1 - e^{-\lambda_i \pi(x^2-1)(r_1^i)^2}] \\
&\overset{(b)}{=} 1 - \frac{1}{x^2},
\end{aligned}
\tag{C.19}
$$

where (a) follows from the Strong Markov property of a stationary PPP which states that conditioned on r_1 of a PPP ϕ, $\phi|_{B(0,r_1)^c}$ is a Poisson point process with the same intensity as ϕ. The equality in (b) follows from the fact that r_1^2 of a stationary PPP of intensity λ is an exponential random variable with mean $\frac{1}{\lambda\pi}$. $\square$

C.0.9 Proof of Corollary 42

Proof. In employing Theorem 39, we need to compute $\mathbb{E}[p_i(\mathrm{SINR}_0^{i,j_i})|\mathbf{r}]$ as follows

$$\mathbb{E}[p_i(\mathrm{SINR}_0^{i,j_i})|\mathbf{r}] = \mathbb{P}\left[\frac{P_i h_1 l(r_1^i)}{N_0 + \sum_{z \geq 2} P_i h_z l(r_z^i)} \geq \beta \left| \frac{r_2^i}{r_1^i} \right.\right]$$

$$= \int_{u=0}^{\infty} \left(c_p\left(1; [u, ut], \lambda_i, P_i, \beta\right) g_{r_1^i | \frac{r_2^i}{r_1^i}}(u, t) du\right), \qquad (\text{C.20})$$

where $c_p\left(1; [u, ut], \lambda_i, P_i, \beta\right)$ is computed through (4.14) and the conditional pdf $g_{r_1^i | \frac{r_2^i}{r_1^i}}(u, t)$ is the distribution of r_1^i conditioned on the ratio $\frac{r_2^i}{r_1^i} = t$.

$$g_{r_1^i | \frac{r_2^i}{r_1^i}}(u, t) = \frac{g_{r_1^i, \frac{r_2^i}{r_1^i}}(u, t)}{\int_{u=0}^{\infty} g_{r_1^i, \frac{r_2^i}{r_1^i}}(u, t) du} \overset{(a)}{=} \frac{g_{r_1^i, \frac{r_2^i}{r_1^i}}(u, t)}{f_{\pi_i}(t)}$$

$$\overset{(b)}{=} 2(\pi \lambda_i)^2 u^3 t^4 e^{-\lambda \pi (ut)^2}, \qquad (\text{C.21})$$

where (a) follows from the fact that the observation is the ratio $\frac{r_2^i}{r_1^i}$ and hence the marginal the pdf $f_{\pi_i}(\cdot)$, which is the derivative of $F_{\pi_i}(\cdot)$ given in Lemma 4.6.2. We now show that (b) holds.

Let the function $g_{r_1, r_1/r_2}(x, y) = (2\pi \lambda_i)^2 xy e^{-\pi \lambda_i y^2}$ denote the joint probability density function for the distance from the origin to the nearest BS and the ratio of distances of the nearest and the second-nearest BSs distributed as a PPP of intensity λ. Transforming this pdf through $(x, y) \to (x, \frac{y}{x})$ yields the joint pdf of $g_{r_1^i, r_2^i/r_1^i}(u, v) = (2\pi \lambda_i)^2 u^3 v e^{-\lambda \pi (uv)^2}$. Plugging the law $F_{\pi_i}(\cdot)$ of Lemma 4.6.2 into (4.8) of Theorem 39 finally completes the proof. $\square$

C.0.10 Proof of Theorem 44

Proof. We start by rearranging (4.8) for our special case where the observations are scalar and the performance $p_i(\cdot)$ and association $\pi_i(\cdot)$ are the same for all technologies i and are independent of i.

$$\mathcal{R}_I^\pi = \sum_{i=1}^T \int_{t \geq 1} \mathbb{E}\left[p(\text{SINR}_0^{i,1}) \middle| \frac{r_2^i}{r_1^i} = t\right] f_{\pi_i}(t) \prod_{j \neq i} F_{\pi_j}(t) dt$$

$$= \sum_{i=1}^T \int_{t \geq 1} \mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i \middle| \frac{r_2^i}{r_1^i} = t\right] f_{\pi_i}(t) \prod_{j \neq i} F_{\pi_j}(t) dt$$

$$\stackrel{(a)}{=} \sum_{i=1}^T \mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i, \frac{r_2^i}{r_1^i} \geq \max_{j \neq i} \frac{r_2^j}{r_1^j}\right]$$

$$= \sum_{i=1}^T \int_{t \geq 1} \mathbb{E}\left[\mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i, \frac{r_2^i}{r_1^i} \geq t \middle| \max_{j \neq i} \frac{r_2^j}{r_1^j} = t\right]\right]$$

where in step (a) we used the fact that the observations from the different technologies are independent. Now using the density function of the maximum of $T - 1$ independent scalar observations each distributed according to a law as given in Lemma 4.6.2, we can simplify the above equation to obtain

$$\mathcal{R}_I^\pi = \begin{cases} \sum_{i=1}^T \int_{x \in [0,1]} \mathbb{P}[\text{SINR}_0^{i,1} \geq \beta_i, \frac{r_2^i}{r_1^i} \geq \frac{1}{x}] \\ 2(T-1)x(1-x^2)^{T-2}dx, & \text{if } T \geq 2 \\ \int_{x \in [0,1]} \mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i, \frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right] dx, & \text{if } T = 1 \end{cases}$$

Further notice that

$$= \mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i, \frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right]$$

$$= \mathbb{P}\left[\text{SINR}_0^{i,j_i} \geq \beta_i \middle| \frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right] \mathbb{P}\left[\frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right]$$

$$= \mathbb{P}\left[\text{SINR}_0^{i,1} \geq \beta_i \middle| \frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right] x^2, \tag{C.22}$$

254

where (C.22) follows from Lemma 4.6.2. Now it remains to show the following lemma, which finally proves (4.20).

Lemma C.0.2. *Assume $N_0^i = 0$ and $l_i(r) = r^{-\alpha}$. For any technology i with intensity λ_i,*

$$\mathbb{P}\left[SIR_0^{i,1} \geq \beta_i \,\middle|\, \frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right] = \frac{1}{1 + \beta_i^{2/\alpha}\phi(\alpha, \beta_i, x)}.$$

In the meantime, akin to the derivation from (4.9) to (4.10), we can obtain the following average achievable rate expression:

$$\mathbb{E}\left[\log_2\left(1 + SIR_0^{i,1}\right)\middle|\,\frac{r_2^i}{r_1^i} \geq \frac{1}{x}\right]$$
$$= \int_{t\geq 0} \frac{1}{1 + (2^t - 1)^{2/\alpha}\phi(\alpha, 2^t - 1, x)}dt.$$

After manipulating this equation in a similar way, we can derive (4.21). $\quad\square$

C.0.11   Proof of Lemma C.0.2

Proof. We drop the subscripts and superscripts i denoting technologies for brevity.

$$\mathbb{P}\left[H_1(r_1)^{-\alpha} \geq \beta \sum_{j\geq 2} H_j(r_j)^{-\alpha} \middle| \frac{r_2}{r_1} \geq \frac{1}{x}\right]$$

$$= \mathbb{E}\left[\mathbb{P}\left[H_1(r_1)^{-\alpha} \geq \beta \sum_{j\geq 2} H_j(r_j)^{-\alpha} \middle| \frac{r_2}{r_1} \geq \frac{1}{x}, r_1\right]\right]$$

$$= \int_{u=0}^{\infty} \mathbb{P}\left[H_1(u)^{-\alpha} \geq \beta \sum_{j\geq 2} H_j(r_j)^{-\alpha} \middle| \frac{r_2}{r_1} \geq \frac{1}{x}, r_1 = u\right]$$

$$f_{r_1 | \frac{r_2}{r_1} \geq \frac{1}{x}}(u)du$$

$$\stackrel{(a)}{=} \int_{u\geq 0} \mathbb{E}\left[\prod_{x\in\phi_i, ||x_i||\geq u/x} \frac{1}{1 + \beta u^\alpha ||x_i||^{-\alpha}}\right]$$

$$\frac{1}{\mathbb{P}\left[\frac{r_2}{r_1} \geq \frac{1}{x}\right]} \int_{v\geq 1/x} f_{r_1, \frac{r_2}{r_1}}(u,v)dv$$

$$\stackrel{(b)}{=} \int_{u\geq 0} \exp\left(-2\pi\lambda \int_{w\geq u/x}\left(1 - \frac{1}{1 + \beta u^\alpha w^{-\alpha}}\right)wdw\right)$$

$$\frac{1}{x^2} 2\lambda\pi u e^{-\lambda\pi(u/x)^2} du$$

$$\stackrel{(c)}{=} \int_{u\geq 0} \exp\left(-\pi\lambda u^2 \beta^{\frac{2}{\alpha}} \int_{b\geq \frac{\beta^{\frac{2}{\alpha}}}{x^2}} \frac{1}{1 + b^{\frac{\alpha}{2}}}db\right) \frac{1}{x^2} 2\lambda\pi u e^{-\lambda\pi(u/x)^2} du$$

$$\stackrel{(d)}{=} \int_{u\geq 0} \exp\left(-\pi\lambda u^2 \beta^{\frac{2}{\alpha}} \frac{1}{x^2} \int_{c\geq \beta^{\frac{-2}{\alpha}}} \frac{1}{1 + x^{-\alpha} c^{\frac{\alpha}{2}}}dc\right)$$

$$\frac{1}{x^2} 2\lambda\pi u e^{-\lambda\pi(u/x)^2} du$$

$$= \frac{1}{1 + \beta^{2/\alpha}\phi(\alpha, \beta, x)}$$

where step (a) follows from the fact that $\{H_i\}$ are i.i.d. exponential random variables as in the proof of Theorem 40. Step (b) follows from Lemmas 4.6.2

and the PGFL of a PPP. Step (c) follows by the substitution $\beta^{\frac{-1}{\alpha}} u^2 b^2 = w$. Step (d) follows by the substitution $x^2 b = c$.

$\square$

Appendix D

Proofs from Chapter 5

D.1 Analysis and Proof of the GBG Algorithm in the Sparse Regime

The following theorem is the main theoretical guarantee on the performance of the GBG algorithm.

Theorem 112. Let $\epsilon \in \left(0, \frac{1}{2}\right)$ be arbitrarily set in Algorithm 2. Let $\eta \in \left(0, \frac{1}{2}\right)$ be such that $\left(\frac{1}{2} + \eta\right)(1 - \epsilon) > \frac{1}{2}$. Then there exists a constant $\lambda_0 < \infty$ depending on $f_{in}(\cdot), f_{out}(\cdot), d, \epsilon$ and η such that for all $\lambda > \lambda_0$, Algorithm 3 will solve weak-recovery.

To prove the main result, we will need an additional classification of the cells of B_n as either T-Good or T-Bad. The nomenclature stands for *Truth*-Good.

Definition 113. A cell Q_z is **T-Good** if -

1. $\left| \phi \cap Q_z \right| \geq \max \left(\lambda \left(\frac{R}{4}\right)^d \frac{1}{d}(1 - \epsilon), 1 \right)$; and

2. For all $i, j \in \mathbb{N}$ such that $X_i, X_j \in Q_{\mathbf{L}_1(z)} \cap \phi$, Pairwise-Classify$(i, j, \phi, G)$ returns $\mathbf{1}_{Z_i = Z_j} - \mathbf{1}_{Z_i \neq Z_j}$, i.e. the ground truth.

If a cell is not T-Good, we call it **T-Bad**.

A cell is T-Good, if for any pair nodes which either lie in the cell under consideration or the neighboring cells, the output of the pairwise estimation matches the ground-truth. Of-course since the ground truth is unknown, one cannot test whether a cell is T-Good or not. We introduce the notion of a T-Good cell to aid in the analysis.

The proof of Theorem 112 can be split into three parts. The first part is composed of combinatorial arguments leveraging the definitions of A-Good and T-Good cells. These combinatorial lemmas will conclude that it suffices to ensure that there exists a 'giant' T-Good connected component in the data (ϕ_n, G_n). The next is a local analysis wherein we conclude that the probability a cell is T-Good can be made arbitrarily large by choosing the constant λ sufficiently high. The final step is to couple the process of T-Good cells to that of dependent site percolation on $\mathbb{Z}^d$ to conclude that if a single cell is T-Good with sufficiently high probability, then there exists a giant T-Good component comprising of many nodes.

D.1.1   Combinatorial Analysis

The main result in this sub-section we want to establish is the following statement. If we establish this, then the performance of our algorithm will follow from a study of the properties of the random graph G.

Proposition 114. If there exists a connected component of T-Good cells in the data (ϕ_n, G_n) which contains a fraction of nodes of G_n strictly larger than

a half with probability $1 - o_n(1)$, then the output returned by Algorithm 3 solves the weak-recovery problem as given in Definition 46.

The proof of the above proposition is based on the following two elementary combinatorial propositions.

Proposition 115. If a cell Q_z is **T-Good**, then it is also **A-Good**. In particular, every connected **T-Good** component is contained in some connected **A-Good** component.

Proof. It suffices to prove that for any $k \in \mathbb{N}$, $\prod_{i=1}^{k}(\mathbf{1}_{Z_i=Z_{i-1}} - \mathbf{1}_{Z_i \neq Z_{i-1}}) = 1$, where $X_k := X_0$ and $Z_k := Z_0$, i.e. a cycle. We can see this by contradiction. Assume $\prod_{i=1}^{k}(\mathbf{1}_{Z_i=Z_{i-1}} - \mathbf{1}_{Z_i \neq Z_{i-1}}) = -1$. This implies that an odd number of $-1's$ exists in the product. This can never be, since this would imply that Z_0 must be both simultaneously $+1$ and -1. Hence, such a product is always $+1$. $\qquad\square$

The following proposition is the basis of Line 5 in the `GBG` in Algorithm 3. For every $z \in \mathbb{Z}^d$, denote by $\mathcal{A}(z)$ the *maximal* $\mathbb{Z}^d$ connected set containing z such that for all $u \in \mathcal{A}(z)$, cell u is A-Good.

Proposition 116. For every $z \in \mathbb{Z}^d$ such that cell z is A-Good, there exists a *unique* partition of $\phi_{\mathcal{A}(z)} := \phi_{\mathcal{A}(z)}^{(+)} \coprod \phi_{\mathcal{A}(z)}^{(-)}$ such that for all $z, z' \in \mathcal{A}(z)$ with $||z - z'||_\infty \leq 1$ and all $X_i \in \phi \cap Q_z$ and $X_j \neq X_i \in \phi \cap Q_{z'}$, we have

- If $X_i \in \phi_z^{(+)}$ and $X_j \in \phi_z^{(-)}$ or , if $X_i \in \phi_z^{(-)}$ and $X_j \in \phi_z^{(+)}$, then Pairwise-Classify(i, j, G) will return -1.

- If $X_i, X_j \in \phi_z^{(+)}$ or if $X_i, X_j \in \phi_z^{(-)}$, then Pairwise-Classify$(i, j, G)$ returns $+1$.

Moreover, the partition produced in Line 8 of our Algorithm 3 coincides with this partition.

This Proposition shows that all nodes inside A-Good connected components can be partitioned into two sets uniquely, such that the T-Good sub-component inside the A-Good component will be partitioned according to the underlying ground truth. Moreover, by following any arbitrary enumeration of the nodes of G as done in Line 5 of Algorithm 3, we can now build this unique partition of nodes of the A-Good component. This is what allows our algorithm to be fast. The proof of this Proposition is quite standard and is defered to the Appendix in **??**. We are now in a position to conclude the proof of Proposition 114.

Proof. Proof of Proposition 114

Proposition 116 justifies Line 5 of Algorithm 3. First note that since every A-Good cell is non-empty of nodes of G, the arbitrary sequence in Line 5 of Algorithm 116 will enumerate all the nodes in each connected component. In other words, the only estimates that will be set in Line 13 of Algorithm 3 are those nodes that fall in the A-Bad cells. Moreover, the partition of the A-Good connected components in Line 8 will coincide with the partition

referred to in Proposition 116. From the definition of T-Good components, the unique partition referred to in Proposition 116 will be such that the T-Good component will be partitioned according to the ground truth. Hence, if there exists a T-Good connected component that has a fraction $\alpha > \frac{1}{2}$ of the nodes of G_n, Algorithm 3 will partition this set of nodes in accordance to the ground truth. Thus, the achieved overlap will be at-least $2\alpha - 1 > 0$. This follows since the mis-classification of all nodes apart from this 'giant' connected T-Good component cannot diminish the overlap below $2\alpha - 1$ which is still positive. $\qquad\square$

D.1.2 Local Analysis

The main goal of this subsection is to show that the probability a cell is T-Good can be made arbitrarily high by taking λ sufficiently high. In order to present the arguments, we recall the definition of a generalized Palm distribution. For any $k \in \mathbb{N}$ and $x_1, \cdots, x_k \in \mathbb{R}^d$, we denote by $\mathbb{P}^{x_1, \cdots x_k}$ to be the Palm distribution of ϕ at $x_1, \cdots x_k$. This measure is the one induced by first sampling ϕ and G and then placing additional points at $x_1, \cdots, x_k$ and equipping them with independent community labels and edges. More precisely, we give these nodes i.i.d. uniform community labels $Z_{-1}, \cdots Z_{-k} \in \{-1, 1\}^k$. Conditionally on all the labels and ϕ, we draw an edge between any $i, j \in \{-k, -(k-1), \cdots\}$ such that at-least one of i or j belong to $\{-k, \cdots, -1\}$ as before, i.e. with probability $f_{in}(||X_i - X_j||)$ if the two nodes have the same community labels or with $f_{out}(||X_i - X_j||)$ if the two nodes have opposite

community labels independently of other edges.

Proposition 117. For any two $x \neq y \in \mathbb{R}^d$ such that $||x - y||_2 < 2R$, then conditionally on the labels of the points at x and y denoted as Z_x and Z_y respectively, we have for all $k \in \mathbb{N}$

- If $Z_x = Z_y$, then $\mathbb{P}^{x,y}[E_G^{(R)}(x,y) = k] = \frac{e^{-\lambda M_{in}(x,y)}(\lambda M_{in}(x,y))^k}{k!}$, i.e. is distributed as a Poisson random variable with mean $\lambda M_{in}(x,y)$.

- If $Z_x \neq Z_y$, then $\mathbb{P}^{x,y}[E_G^{(R)}(x,y) = k] = \frac{e^{-\lambda M_{out}(x,y)}(\lambda M_{out}(x,y))^k}{k!}$, i.e. is distributed as a Poisson random variable with mean $\lambda M_{out}(x,y)$.

Proof. Slivnyak's theorem for independently marked PPP gives that conditionally on k points at locations $x_1, \cdots x_k \in \mathbb{R}^d$, the marked point process $\bar{\phi} \setminus \{x_1, \cdots, x_k\}$ has the same distribution as the original marked point process, i.e. is a PPP of intensity λ with independent marks. The independent thinning property of the PPP states that if any point at $x \in \phi$ is retained with probability $p(x)$ and deleted with probability $1 - p(x)$, independently of everything else, then the set of points not deleted forms a (potentially inhomogeneous) PPP.

Notice that the event that any $k \in \phi \setminus \{x, y\}$ such that $k \in B(x, R) \cap B(y, R)$ has an edge to both points x and y in G only depends on the location k and the community labels of points at locations k, x and y and is independent of everything else. Now, since the community labels are i.i.d. and independent

263

of ϕ, the independent thinning property of PPP gives that the distribution of $E_G^{(R)}(x, y)$ is a Poisson random variable.

It remains to notice that the means are precisely $\lambda M_{in}(x, y)$ and $\lambda M_{out}(x, y)$. This follows from the Campbell - Mecke's theorem, that for any $F(\cdot) : \mathbb{R}^d \to \mathbb{R}_+$, we have for independently marked process is

$$\mathbb{E}_\phi^{x,y}\left[\sum_{z \in \phi \backslash \{x,y\}} F(z)\right] = \lambda \int_{z \in \mathbb{R}^d} \mathbb{E}_\phi^{x,y,z}[F(z)]dz. \tag{D.1}$$

Now, setting $F(z) := \mathbf{1}_{z \text{ has an edge to } x \text{ and } y}\mathbf{1}_{||z-x||_2<R}\mathbf{1}_{||z-y||_2<R}$ will conclude the statement on the means. $\qquad \square$

Proposition 118. For all connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 54, there exists a constant $c > 0$ such that for all $x \neq y \in \mathbb{R}^d$ satisfying $||x - y||_2 \leq (3/4)R$, we have

$$\mathbb{P}^{x,y}[(x, y) \text{ is misclassified by Algorithm 1}] \leq e^{-c\lambda}, \tag{D.2}$$

where the constant c satisfies

$$c \geq \inf_{x,y \in \mathbb{R}^d:||x-y||_2\leq 3R/4}$$
$$(\mathbf{1}_{M_{out}(x,y)>0}M_{out}(x,y) + \mathbf{1}_{M_{out}(x,y)=0}M_{in}(x,y))h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right),$$
$$\tag{D.3}$$

where $h(t) := (1 + t)\log(1 + t) - t$, for all $t \in \mathbb{R}_+$. In particular, $c > 0$ is strictly positive.

Proof. From Proposition 117, we know that $E_G^{(R)}(x, y)$ is either a Poisson random variable with mean $\lambda M_{in}(x, y)$ if the two nodes have the same community label or is a Poisson random variable of mean $\lambda M_{out}(x, y)$ if the two nodes have opposite community labels. Thus, the probability of mis-classification is then

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] =$$
$$\frac{1}{2}\mathbb{P}\left[X \geq \lambda\frac{M_{in}(x, y) + M_{out}(x, y)}{2}\right] + \frac{1}{2}\mathbb{P}\left[Y \leq \lambda\frac{M_{in}(x, y) + M_{out}(x, y)}{2}\right],$$
$$\text{(D.4)}$$

where X is a Poisson random variable of mean $\lambda M_{out}(x, y)$ and Y is a Poisson random variable of mean $\lambda M_{in}(x, y)$. The above interpretation is a probabilistic restatement of Algorithm 1. The coefficient $1/2$ denotes the case that the points at x and y could be in the same community or in opposite communities. Thus, by a basic application of Chernoff's bound, we have

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq$$
$$\frac{1}{2}e^{-\lambda M_{out}(x,y)h\left(\frac{M_{in}(x,y)-M_{out}(x,y)}{2M_{in}(x,y)}\right)} + \frac{1}{2}e^{-\lambda M_{in}(x,y)h\left(\frac{M_{in}(x,y)-M_{out}(x,y)}{2M_{in}(x,y)}\right)}, \quad \text{(D.5)}$$

where $h(\cdot)$ is defined in the statement of the proposition.

Now under the assumptions on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$, for all $r \in [\tilde{r}, R]$, $f_{in}(r) > f_{out}(r)$, we have that, $\inf_{x,y\in\mathbb{R}^d:||x-y||_2\leq(3/4)R} M_{in}(x, y) - M_{out}(x, y) > 0$. Moreover, since $M_{in}(x, y)$ and $M_{out}(x, y)$ are non-negative, $M_{in}(x, y) - M_{out}(x, y) > 0$ implies automatically that $M_{in}(x, y) > 0$ for all $x, y \in \mathbb{R}^d$ such that $||x - y||_2 \leq (3R/4)$. Hence, it

follows that

$$\sup_{x,y \in \mathbb{R}^d : ||x-y||_2 \leq (3/4)R} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq e^{-c\lambda}, \quad (D.6)$$

where c is a strictly positive constant as given in the statement of the proposition.

$\square$

Lemma 119. For all $z \in \mathbb{Z}^d$,

$$\mathbb{P}[\text{Cell } z \text{ is T-Good in graph } G] \geq 1 - e^{-\lambda(R/4)^d \frac{1}{d}h(\epsilon)} - \lambda^2(3R/4)^d\frac{1}{d}e^{-c\lambda}, \tag{D.7}$$

where the constant c and function $h(\cdot)$ are defined in Proposition 118.

Proof. This follows from a basic union bound. We will prove an upper bound to a cell being T-Bad. A cell is T-Bad if either the number of points is smaller than $\lambda(R/4d^{1/d})^d(1 - \epsilon)$ or there exists two points X_i and X_j in the 1 thickening of the cell $\{z\}$ such that when Algorithm 1 is run on input (i, j, G), the returned answer is different from the truth.

From a simple Chernoff bound, the probability that a cell has fewer than $\lambda(R/4d^{1/d})^d(1 - \epsilon)$ is at-most $e^{-\lambda(R/4d^{1/d})^d h(\epsilon)}$, where $h(\epsilon)$ is strictly positive for all $\epsilon > 0$.

We bound the probability that there exist two nodes that Algorithm 1 mis-classifies by the first moment method. We use the fact that if $X \geq 0$ is a $\mathbb{N}$ valued random variable, then $\mathbb{P}[X > 0] \leq \mathbb{E}[X]$. Hence, the probability

that there exists a pair of points of ϕ that are mis-classified is bounded by the average number of pairs of points that are misclassified. Thus, for each cell z, we compute

$$\mathbb{E}[\sum_{i,j\in\mathbb{N}} \mathbf{1}_{X_i,X_j\in\mathbf{L}_1(z)} \mathbf{1}_{\text{Algorithm 1 mis-classifies } i \text{ and } j}]. \tag{D.8}$$

From the Moment-Measure expansion and the Campbell-Mecke theorem for an independently marked PPP ([119]), we obtain

$$\mathbb{E}[\sum_{i,j\in\mathbb{N}} \mathbf{1}_{X_i,X_j\in\mathbf{L}_1(z)} \mathbf{1}_{\text{Algorithm 1 mis-classifies } i \text{ and } j}]$$
$$= \lambda^2 \int_{x\in Q_{\mathbf{L}_1(z)}} \int_{y\in Q_{\mathbf{L}_1(z)}} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }]dxdy$$
$$\leq \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \tag{D.9}$$

The last inequality follows directly from Proposition 118. Therefore, by a simple union bound, we see that

$$\mathbb{P}[\text{ Cell } z \text{ is T-Bad}] \leq e^{-\lambda(R/4d^{1/d})^d h(\epsilon)} + \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \tag{D.10}$$

The proposition is proved by taking complements. $\qquad\square$

Thus, we immediately have the following corollary which is what we will use in the sequel. The key fact to be used here is that the tessellation size R does not depend on λ and only depends on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$.

Corollary 120. For every $p \in (0,1)$, and every $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 54, there exists a λ' such that for all $\lambda > \lambda'$, and all $z \in \mathbb{Z}^d$, $\mathbb{P}[\text{Cell } z \text{ is T-Good}] \geq p$.

Proof. It suffices to notice that for each fixed $f_{in}(\cdot)$, $f_{out}(\cdot)$ and d, we have

$$\lim_{\lambda \to \infty} p(\lambda) \geq \lim_{\lambda \to \infty} 1 - e^{-\lambda (R/4)^d \frac{1}{d} h(\epsilon)} - (\lambda^2 + \lambda)(3R/4)^d \frac{1}{d} e^{-c\lambda} = 1, \quad \text{(D.11)}$$

where c is given in Proposition 118.

$\square$

D.1.3 Global Analysis

In this section, we present the central tool required to analyze about the 'giant' connected T-Good component in the graph G_n. To do so, we exploit a coupling between the T-Good cells in the graph G and a certain dependent site percolation process on $\mathbb{Z}^d$.

Denote by $(Y_z)_{z \in \mathbb{Z}^d}$ to be the random $0 - 1$ field on $\mathbb{Z}^d$ where $Y_z := \mathbf{1}_{\text{Cell } z \text{ is T-Good in } G}$. From the construction of the field, notice that the random field $(Y_z)_{z \in \mathbb{Z}^d}$ is only mildly dependent. Indeed, given any two $z, z' \in \mathbb{Z}^d$, such that $||z - z'||_1 \geq 12 d^{1/d}$, we have that Y_z and $Y_{z'}$ are independent random variables. This follows from the fact that we only look upto Euclidean distance of at-most $2R$ from any point inside a cell z to determine whether a cell is T-Good or T-Bad. Since, in an independently marked PPP, events corresponding to disjoint sets of $\mathbb{R}^d$ are independent, the claim follows.

We will now set some notation that will be useful in studying the process $(Y_z)_{z \in \mathbb{Z}^d}$. For any $z \in \mathbb{Z}^d$, cell z is **open in** $\mathbb{Z}^d$ if $Y_z = 1$. Similarly, any edge connecting z and z' is said to be open if both its end points are open. For any $z \in \mathbb{Z}^d$, we denote by $\mathcal{C}(z)$ to be the maximal connected random subset of $\mathbb{Z}^d$ containing z such that all $z' \in \mathcal{C}(z)$ satisfies $Y_{z'} = 1$. The main proposition we want to establish in this section is the following.

Proposition 121. For every $\eta \in \left(0, \frac{1}{2}\right)$, there exists $\lambda_0(\eta, \epsilon) < \infty$ (where ϵ is set in Algorithm 2) chosen sufficiently high (as a function of $f_{in}(r), f_{out}(r), r \in [0, R]$ and d), such that for all $\lambda > \lambda_0(\eta, \epsilon)$ and all $j \in \mathbb{Z}^d$

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i - j||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(j)} \geq \frac{1}{2} + \eta, \tag{D.12}$$

$\mathbb{P}$ almost-surely on the event that $\{|\mathcal{C}(j)| = \infty\}$. Moreover, for $\lambda > \lambda_0(\eta, \epsilon)$, and all $j \in \mathbb{Z}^d$, $\mathbb{P}[|\mathcal{C}(j)| = \infty] \geq \frac{1}{2} + \eta$ and $\mathbb{P}[\exists j \in \mathbb{Z}^d : |\mathcal{C}(j)| = \infty] = 1$.

The key insight out of the proposition we want is to ensure that by taking λ sufficiently high, there exists an infinite open component in the process $(Y_z)_{z \in \mathbb{Z}^d}$, i.e. there exists $z \in \mathbb{Z}^d$ such that $|\mathcal{C}(z)| = \infty$. Moreover, we want to show that this infinite component contains more than half of the sites of $\mathbb{Z}^d$. The reason this does not immediately follow from Corollary 120 is that we have not yet established that the infinite open component in $(Y_z)_{z \in \mathbb{Z}^d}$ if it exists is unique. However [242] provides a clean 'black-box' methodology to establish this and our proposition can be viewed as a direct corollary of Theorem 1 in [242]. We will first dominate the process $(Y_z)_{z \in \mathbb{Z}^d}$ by an independent

269

percolation process which is known to have a unique infinite component and then leverage this domination to conclude the proposition.

Proof. Notice that the process $(Y_z)_{z \in \mathbb{Z}^d}$ is $M := \lceil 12 d^{1/d} \rceil$ dependent. Moreover, thanks to Proposition 118, for every $z \in \mathbb{Z}^d$,

$$\mathbb{P}[Y_z = 1 | \sigma(Y_u : u \in \mathbb{Z}^d, ||u - z||_\infty > M)] \geq p(\lambda), \ \mathbb{P} \text{ a.s. }, \quad \text{(D.13)}$$

where $p(\lambda) \to 1$ as $\lambda \to \infty$.

Thus, from Theorem 1 in [242], the law of $(Y_z)_{z \in \mathbb{Z}^d}$ stochastically dominates that of i.i.d. Bernoulli $\tilde{p}(\lambda)$ random variables where $\tilde{p}(\lambda)$ converges to 1 as $p(\lambda)$ converges to 1. More precisely, Theorem 1 from [242] gives the existence of a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ containing two sequences of $\{0, 1\}$ valued random variables $(Y_z')_{z \in \mathbb{Z}^d}$ and $(\tilde{Y}_z')_{z \in \mathbb{Z}^d}$ such that

- The distribution of $(Y_z')_{z \in \mathbb{Z}^d}$ is the same as that of $(Y_z)_{z \in \mathbb{Z}^d}$.

- For all $z \in \mathbb{Z}^d$, $Y_z' \geq \tilde{Y}_z'$, $\mathbb{P}'$ almost-surely.

- $\mathbb{P}'[\tilde{Y}_z' = 1 | \sigma(\tilde{Y}_u' : u \in \mathbb{Z}^d \setminus \{z\})] = \tilde{p}(\lambda)$, $\mathbb{P}'$ almost-surely. In other words, $(\tilde{Y}_z')_{z \in \mathbb{Z}^d}$ is an i.i.d. sequence of Bernoulli random variables with success probability $\tilde{p}(\lambda)$.

- $\tilde{p}(\lambda)$ converges to 1 as $p(\lambda)$ converges to 1.

Denote by $\mathcal{C}'(0)$ and $\tilde{C}'(0)$ the cluster at the origin of the process $(Y_z')_{z \in \mathbb{Z}^d}$ and $(\tilde{Y}_z')_{z \in \mathbb{Z}^d}$ respectively. Denote by $\theta_d(\lambda) := \mathbb{P}'[|\tilde{C}'(0)| = \infty]$. From

270

a direct application of Peirl's argument ([83], Chapter 1), it is also well know that $\theta_d(\lambda) \to 1$ as $\tilde{p}(\lambda) \to 1$. Thanks to Line 4 above, we have $\theta_d(\lambda) \to 1$ as $p(\lambda) \to 1$. From Corollary 120, this can be rephrased as $\lim_{\lambda \to \infty} \theta_d(\lambda) = 1$.

The stochastic domination in Line 2 above yields

$$\frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}'(j)} \geq \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i \in \tilde{\mathcal{C}}'(j)} \quad \mathbb{P}' \text{ a.s.} \qquad \text{(D.14)}$$

On the event that $|\tilde{\mathcal{C}}'(j)| = \infty$, we have

$$\frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}'(j)} \geq \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}'(i)|=\infty} \quad \mathbb{P}' \text{ a.s.} \qquad \text{(D.15)}$$

This follows from the well known fact that in an independent site percolation process that the infinite component if it exists is unique. In other-words, for all $i, j \in \mathbb{Z}^d$, $|\tilde{\mathcal{C}}'(i)| = \infty$ and $|\tilde{\mathcal{C}}'(j)| = \infty$ implies $\tilde{\mathcal{C}}'(i) = \tilde{\mathcal{C}}'(j)$, $\mathbb{P}'$ almost-surely. Now, taking a limit on both sides, we get that

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}'(j)} \geq \liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}'(i)|=\infty} \quad \mathbb{P}' \text{ a.s.}$$
$$\text{(D.16)}$$

From Birkhoff's ergodic theorem, it is well known that for all $j \in \mathbb{Z}^d$,

$$\lim_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}'(i)|=\infty} = \theta_d(\lambda) \quad \mathbb{P}' \text{ a.s.} \qquad \text{(D.17)}$$

But since $\lim_{\lambda \to \infty} \theta_d(\lambda) = 1$, for every η and ϵ, we can take $\lambda_0(\eta, \epsilon)$ sufficiently large so that $p(\lambda)$ is sufficiently large which in turn indicates $\tilde{p}(\lambda)$ is sufficiently large so that $\theta_d(\lambda) \geq \frac{1}{2} + \eta$. The proof is concluded by observing that $(Y'_z)_{z \in \mathbb{Z}^d} \overset{(d)}{=} (Y_z)_{z \in \mathbb{Z}^d}$.

$\square$

D.1.4 Concluding that Weak-Recovery is Solvable

The following proposition along with Proposition 114 will conclude the proof of Theorem 112.

Proposition 122. Let $\epsilon \in \left(0, \frac{1}{2}\right)$ be set in Algorithm 2. For all $\eta \in \left(0, \frac{1}{2}\right)$ such that $(1 - \epsilon)\left(\frac{1}{2} + \eta\right) > \frac{1}{2}$, for all $\lambda \geq \lambda_0(\epsilon, \eta)$ where $\lambda_0(\epsilon, \eta)$ is from Proposition 121, the fraction of nodes of G_n that lie in the largest T-Good component, denoted by $\alpha_n \in [0, 1]$ is such that $\liminf_{n \to \infty} \alpha_n > \frac{1}{2}$, $\mathbb{P}$ almost-surely.

Proof. Observe that the definition of a cell being A-Good or A-Bad is spatially 'local'. More precisely, for all $z \in \mathbb{Z}^d$ such that $z + B(0, 2R) \in B_n$, the event that cell z being A-Good in G_n is the same as cell being A-Good in G. We call cells $z \in \mathbb{Z}^d$ such that $z + B(0, 2R) \in B_n$ *internal* to B_n. Observe that all $z \in \mathbb{Z}^d$ is eventually internal to B_n for all n large enough. Moreover, since each cell is of side $R/(4d^{1/d})$, B_n has at-most $\lceil (4n^{1/d}/Rd^{1/d})^d \rceil$ cells out-of which at-least $\lfloor (4n^{1/d}/Rd^{1/d})^d \rfloor - \lceil 8dn^{1/d} \rceil$ cells are 'internal' to B_n. Thus, the fraction of cells in B_n that are internal to B_n is $1 - o_n(1)$.

From Proposition 121, we know that $\mathbb{P}[|\mathcal{C}(0)| = \infty] \geq \frac{1}{2} + \eta$ and $\mathbb{P}[\exists z \in \mathbb{Z}^d : |\mathcal{C}(z)| = \infty] = 1$. Moreover on the event $\{|\mathcal{C}(z)| = \infty\}$, we know from Proposition 121 that

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i - z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \; \mathbb{P} \text{ a.s.} \qquad (D.18)$$

However, from an elementary counting argument, we conclude that

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \; \mathbb{P} \text{ a.s.} \qquad (D.19)$$

In other words, the reference point does not matter when considering the limit, which can be seen easily by the following in Equation (D.20)

.

$$\frac{1}{(2(n+z))^d} \sum_{i\in\mathbb{Z}^d:||i-z||_\infty\leq n} \mathbf{1}_{i\in\mathcal{C}(z)} \leq \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d:||i-z||_\infty\leq n} \mathbf{1}_{i\in\mathcal{C}(z)}$$

$$\leq \frac{1}{(2(n+z))^d} \left(\sum_{i\in\mathbb{Z}^d} \mathbf{1}_{||i-z||_\infty\geq n}\mathbf{1}_{||i||_\infty\leq z+n} + \mathbf{1}_{||i-z||_\infty\leq n}\mathbf{1}_{i\in\mathcal{C}(z)} \right).$$

But since for every fixed $z \in \mathbb{Z}^d$,

$$\frac{1}{(2(n+z))^d} \left(\sum_{i\in\mathbb{Z}^d} \mathbf{1}_{||i-z||_\infty\geq n}\mathbf{1}_{||i||_\infty\leq z+n} + \mathbf{1}_{||i-z||_\infty\leq n}\mathbf{1}_{i\in\mathcal{C}(z)} \right) -$$

$$\frac{1}{(2(n+z))^d} \sum_{i\in\mathbb{Z}^d:||i-z||_\infty\leq n} \mathbf{1}_{i\in\mathcal{C}(z)} = O\left(n^{1-d}\right),$$

it follows that for all $z \in \mathbb{Z}^d$

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d:||i-z||_\infty\leq n} \mathbf{1}_{i\in\mathcal{C}(z)} = \liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d:||i||_\infty\leq n} \mathbf{1}_{i\in\mathcal{C}(z)}. \quad \text{(D.20)}$$

Let $z \in \mathbb{Z}^d$ be arbitrary and condition on the event $\{|\mathcal{C}(z)| = \infty\}$. On this event, Equation (D.20) and Proposition 121 along with the fact that the fraction of cells in B_n that are internal is $1 - o_n(1)$ give that the fraction of internal cells in B_n in the connected T-Good component of cell z (i.e. in $\mathcal{C}(z)$) is $\frac{1}{2}+\eta-o_n(1)$. Since there are at-least $\lambda(R/4)^d(1/d)(1-\epsilon)$ nodes of G in each T-Good cell, the number of nodes of G_n in this T-Good connected component is at-least $\left(\lfloor(4n^{1/d}/R)^d\rfloor - \lceil 8dn^{1/d}\rceil\right)\left(\frac{1}{2} + \eta\right)\lambda(R/4)^d(1-\epsilon) > \frac{1}{2}(\lfloor(4n^{1/d}/R)^d\rfloor - \lceil 8dn^{1/d}\rceil)\lambda(R/4)^d$ since we assumed that $(1 - \epsilon)\left(\frac{1}{2} + \eta\right) > \frac{1}{2}$. Moreover, from

273

elementary Chernoff and Borell Cantelli arguments, we get that for every fixed $\epsilon' > 0$, there exists a random $n_{\epsilon'}$ such that for all $n \geq n_{\epsilon'}$, the number of nodes in G_n is less than or equal to $\lceil (4n^{1/d}/R)^d \rceil \lambda(R/4)^d(1+\epsilon')$ almost-surely. Now, fix an $\epsilon' > 0$, such that there exists a $\gamma > 0$ satisfying $\frac{(1/2+\eta)(1-\epsilon)}{(1+\epsilon')} = \frac{1}{2} + \gamma$. Thus, for n larger than $n_{\epsilon'}$, the fraction of nodes in G_n lying the T-Good component of cell z is α_n, where

$$\alpha_n \geq \frac{\left(\lfloor (4n^{1/d}/R)^d \rfloor - \lceil 8dn^{1/d} \rceil\right)\left(\frac{1}{2}+\eta\right)\lambda(R/4)^d(1-\epsilon)}{\lceil (4n^{1/d}/R)^d \rceil \lambda(R/4)^d(1+\epsilon')} \geq \frac{1}{2} + \gamma - o_n(1)$$

(D.21)

almost-surely, i.e., $\lim_{n\to\infty} \mathbb{P}\left[\alpha_n > \frac{1}{2} \middle| |\mathcal{C}(z)| = \infty\right] = 1$. But since $\mathbb{P}[\exists z \in \mathbb{Z}^d : |\mathcal{C}(z)| = \infty] = 1$, we can drop the conditioning on the event $\{|\mathcal{C}(z)| = \infty\}$ and conclude that with probability 1, a fraction of nodes of G_n strictly larger than half lie in a connected T-Good component.

$\square$

D.1.5   Proof of Proposition 55

From Proposition 122, we know that for every $\epsilon \in (0,1)$ and $\eta \in (0, \frac{1}{2}]$), there exists $\lambda_0(\epsilon, \eta) < \infty$, such that for all $\lambda > \lambda_0(\epsilon, \eta)$, the GBG algorithm achieves an overlap of $(\frac{1}{2} + \eta)(1 - \epsilon)$. The proof is concluded by noticing that for any $\delta \in (\frac{1}{2}, 1)$, we can choose $\epsilon \in (0,1)$ and $\eta \in (\frac{1}{2}, 1)$ such that $(\frac{1}{2} + \eta)(1 - \epsilon) > \delta$.

D.2 Identifiability of the Partition and Proof of Theorem 56

The key technical tool is the ergodicity of the PPP which is summarized in the following lemma. We need to set some notation that are needed to state the lemma. Denote by $\mathbb{M}_\Xi(\mathbb{R}^d)$ the set of all 'marked' point processes on $\mathbb{R}^d$ where each point is assigned a 'mark' from the measure space Ξ, with its associated sigma-algebra. The set $\mathbb{M}_\Xi(\mathbb{R}^d)$ is a Polish space, has a natural topology and hence an associated sigma-algebra (see [136]). Denote by $\theta : \mathbb{R}^d \times \mathbb{M}_\Xi(\mathbb{R}^d) \to \mathbb{M}_\Xi(\mathbb{R}^d)$ the 'shift' operator which is a measurable function where $\theta(x, \psi)$ retains the same marks but translates all points of ψ by a vector x.

Lemma 123. Let $C_n \subset \mathbb{R}^d$ be a sequence of L_p, $p \in [1, \infty]$ balls centered at the origin with radius going to infinity as $n \to \infty$. Let $f : \mathbb{M}_\Xi(\mathbb{R}^d) \to \mathbb{R}_+$ be a measurable function such that $\mathbb{E}^0_\phi[f] < \infty$. Then, the following limit exists :

$$\lim_{n \to \infty} \frac{\sum_{i \in \mathbb{N}} \mathbf{1}_{X_i \in C_n} f \circ \theta(X_i, G)}{\sum_{i \in \mathbb{N}} \mathbf{1}_{X_i \in C_n}} = \mathbb{E}^0[f] \; \mathbb{P} \text{ a.s. } . \tag{D.22}$$

We now prove Theorem 56.

Proof. First if $g(\cdot) \neq \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}$ Lebesgue almost everywhere, then G and $H_{\lambda, g(\cdot), d}$ are mutually singular and this can be seen through the following elementary argument. Fix some $L < \infty$ such that $\int_{x \in \mathbb{R}^d : ||x|| \leq L} g(||x||) dx \neq \int_{x \in \mathbb{R}^d : ||x|| \leq L} ((f_{in}(x) + f_{out}(||x||))/2) dx$ are both finite. Such a L exists since

$g(\cdot) \neq \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}$. Now, we apply the ergodic theorem, where every node $i \in \mathbb{N}$ of ϕ is equipped with a mark $\Xi_i \in \mathbb{N}$, which denotes the number of graph neighbors of node i at a distance of at-most L from X_i, i.e. $\Xi_i = |\{j \in \mathbb{N} \setminus \{i\} i \sim_G j, ||X_i - X_j|| \leq L\}|$. Thus, the Ergodic theorem implies that the measure induced by G will be concentrated on the set

$$\left\{ g \in \mathbb{M}_\mathcal{G}(\mathbb{R}^d) : \lim_{n \to \infty} \frac{\sum_{i,j \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq n, ||X_i - X_j|| \leq L} \mathbf{1}_{i \sim_g j}}{\sum_{i,j \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq n, ||X_i - X_j|| \leq L}} \right.$$
$$= \int_{x \in \mathbb{R}^d : ||x|| \leq L} (f_{in}(x) + f_{out}(||x||))/2 dx \Bigg\},$$

while the measure induced by $H_{\lambda, g(\cdot), d}$ will be concentrated on the set

$$\left\{ g \in \mathbb{M}_\mathcal{G}(\mathbb{R}^d) : \lim_{n \to \infty} \frac{\sum_{i,j \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq n, ||X_i - X_j|| \leq L} \mathbf{1}_{i \sim_g j}}{\sum_{i,j \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq n, ||X_i - X_j|| \leq L}} \right.$$
$$= \int_{x \in \mathbb{R}^d : ||x|| \leq L} g(||x||) dx \Bigg\}.$$

Thus, the only case to consider is the one where $g(\cdot) = (f_{in}(\cdot) + f_{out}(\cdot)/2$ Lebesgue almost-everywhere. From linearity of expectation, the average degree of any node $i \in \mathbb{N}$ in both graphs G and $H_{(\lambda, \frac{f_{in}(\cdot) + f_{out}(\cdot)}{2}, d)}$ is the same and equal to $(\lambda/2) \int_{x \in \mathbb{R}^d} (f_{in}(||x||) + f_{out}(||x||)) dx$ and thus empirical average of the degree does not help. However, we see that the triangle profiles differ in the two models which we leverage to prove the Theorem.

For ease of notation, we denote by $H := H_{\lambda, (f_{in}(\cdot) + f_{out}(\cdot))/2, d}$. Define

$$\Delta = \mathbb{E}^0 \left[\sum_{x \neq y \neq 0} h(x, y) \mathbf{1}((0, x) \in E, (0, y) \in E, (x, y) \in E) \right].$$

276

Denote by Δ_G and Δ_H the value of the above expression if the underlying graphs were G and H respectively. From the moment measure expansion of PPPs [119], we get that

$$\Delta_G = \int_{x \in \mathbb{R}^d} \int_{y \in \mathbb{R}^d} h(x,y)(f_{in}(||x-y||) \left(\frac{f_{in}(||x||)f_{in}(||y||) + f_{out}(||x||)f_{out}(||y||)}{4} \right)$$
$$+ f_{out}(||x-y||) \left(\frac{f_{in}(||x||)f_{out}(||y||) + f_{out}(||x||)f_{in}(||y||)}{4} \right))\lambda^2 dx dy,$$

and

$$\Delta_H = \int_{x \in \mathbb{R}^d} \int_{y \in \mathbb{R}^d} h(x,y) \left(\frac{f_{in}(||x||) + f_{out}(||x||)}{2} \right) \left(\frac{f_{in}(||y||) + f_{out}(||y||)}{2} \right)$$
$$\left(\frac{f_{in}(||x-y||) + f_{out}(||x-y||)}{2} \right) \lambda^2 dx dy.$$

Now observe that

$$\Delta_G - \Delta_H = \int_{x \in \mathbb{R}^d} \int_{y \in \mathbb{R}^d} h(x,y) \frac{1}{2} \left(f_{in}(||x||) - f_{out}(||x||) \right) \left(f_{in}(||y||) - f_{out}(||y||) \right)$$
$$\left(f_{in}(||x-y||) - f_{out}(||x-y||) \right) \lambda^2 dx dy.$$

From the fact that $f_{in}(r) \geq f_{out}(r)$ and $f_{in}(r)$ is not equal to $f_{out}(r)$ Lebesgue almost everywhere, there exists a positive bounded function $h(x,y)$ such that $0 \leq \Delta_H < \Delta_G < \infty$. Choose one such test function $h(\cdot, \cdot)$, for ex. $h(x,y) = \mathbf{1}_{||x|| \leq R} \mathbf{1}_{||y|| \leq R} \mathbf{1}_{f_{in}(||x||) > f_{out}(||x||)} \mathbf{1}_{f_{in}(||y||) > f_{out}(||y||)} \mathbf{1}_{f_{in}(||x-y||) > f_{out}(||x-y||)}$ and consider the following estimator:

Algorithm 8 Detect-Partitions

Given the data, i.e. locations of nodes and the graph, pick a L large enough and compute

$$\Delta^{(L)} := \frac{\sum_{i\in\mathbb{N}}\mathbf{1}(|X_i|\leq L)\tilde{h}(X_i)}{\sum_{i\in\mathbb{N}}\mathbf{1}(|X_i|\leq L)},$$

where $\tilde{h}(X_i) = \sum_{j,k\in\mathbb{N},j\neq k\neq i} h(X_i - X_j, X_i - X_k)\mathbf{1}(i\sim_G j, i\sim_G k, j\sim_G k)$.

From ergodicity, we know that $\lim_{L\to\infty}\Delta^{(L)} = \Delta_G$, $\mathbb{P}$ almost-surely if the data is the block model graph or $\lim_{L\to\infty}\Delta^{(L)} = \Delta_H$ $\mathbb{P}$ almost surely if the graph is drawn according to the null model. We can apply the ergodic theorem since a spatial random graph is a marked point process (as described in Section 5.2.1). Thus, the measures induced by (ϕ, G) and $(\phi, H_{\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d})$ are mutually singular. Moreover, the above algorithm when tested on the finite data (G_n, ϕ_n) runs in time proportional to λn with the multiplicative constants depending on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ with success probability $1 - o_n(1)$ with the $o_n(1)$ term depending on the function $h(\cdot,\cdot)$ chosen.

$\square$

We investigated the singularity of measures in order to understand the question of distinguishability of the planted partition model. This is a hypothesis testing problem of whether the data (graph and spatial locations) is drawn from the distribution of G or from the distribution of $H_{\lambda,g(\cdot),d}$ with a probability of success exceeding a half given a uniform prior over the models. This

problem is in some sense easier than Community Detection, since this asks for asserting whether a partition exists or not, which intuitively should be simpler than finding the partition. Indeed, we show this in our model by proving that the distinguishability problem is trivially solvable while community detection undergoes a phase transition and is solvable only under certain regimes. In the general sparse SBM however, the equivalence between distinguishability and community detection is only conjectured and not yet proven ([142],[62]).

D.3   The Exact-Recovery Problem

In this section, we provide a lower bound for the exact-recovery problem as stated in Definition 47. Recall that for the exact-recovery case, we equipped the set B_n with the torridal metric rather than the usual Euclidean metric. This is done mainly to simplify the presentation of the results. Nevertheless, one could establish identical results to the case when the set B_n equipped with the Euclidean metric albeit with significantly more heavier notation to handle the 'edge effects'. We note that for any $x := (x_1, \cdots, x_d), y := (y_1, \cdots, y_d) \in B_n$, the torroidal distance is given by $||x - y||_{\mathcal{T}_n} = ||(\min(|x_1 - y_1|, n^{1/d} - |x_1 - y_1|), \cdots, \min(|x_d - y_d|, n^{1/d} - |x_d - y_d|))||$, where $||.||$ is the Euclidean norm on $\mathbb{R}^d$. The key result we establish about this model is a lower bound or a necessary condition to perform Exact-Recovery. We then observe that the GBG algorithm presented earlier achieves Exact-Recovery, if the intensity λ is sufficiently high, thereby establishing a phase-transition. We establish the lower bound by exploiting recent advances in the understanding of error

279

exponents for the probability of error in distinguishing between two Poisson random vectors developed in [20]. Furthermore, we also conjecture this necessary condition to also be sufficient. The exact proof of this conjecture is left open in this paper. Another important note is that since we view the set B_n as a torus in this section, the notion of Palm probability needs more care in stating. In the most general sense, we must employ the moment measure expansions of a Poisson Point Process on compact topological groups to discuss the Palm measure on the torus. We provide a justification of the moment measure equations we use in this section in the Appendix D.5.

D.3.1   Lower Bound for Exact-Recovery

In this section, we prove Theorem 58. To do so, we first give a general result in Proposition 124 using the Genie aided argument introduced in [18], and then subsequently prove an explicit formula for the lower bound stated in Theorem 58 using the large deviation results of [20] for hypothesis testing between Poisson random vectors. To state the result, we define a notion of *Flip-Bad* for nodes. Roughly speaking, we say a node $i \in [1, N_n]$ is Flip-Bad in G_n, if on flipping the community label of node i from Z_i to $-Z_i$, the likelihood of the observed data (G_n, ϕ_n) increases. More formally, given data

$(\phi_n, G_n, (Z_i)_{i=1}^{N_n})$, the likelihood is defined as

$$\mathcal{L}(\phi_n, G_n, (Z_i)_{i=1}^{N_n}) := \frac{e^{-\lambda n}(\lambda n)^{N_n}}{N_n!} \left(\frac{1}{2n}\right)^{N_n} \prod_{1 \leq i < j \leq N_n : Z_i = Z_j, i \sim j} f_{in}^{(n)}(||X_i - X_j||)$$

$$\prod_{1 \leq i < j \leq N_n : Z_i \neq Z_j, i \sim j} f_{out}^{(n)}(||X_i - X_j||) \prod_{1 \leq i < j \leq N_n, Z_i = Z_j, i \nsim j} (1 - f_{in}^{(n)}(||X_i - X_j||))$$

$$\prod_{1 \leq i < j \leq N_n, Z_i \neq Z_j, i \nsim j} (1 - f_{out}^{(n)}(||X_i - X_j||)),$$

with the empty product being equal to 1. Thus, it is easy to see that for every n, $0 < \mathcal{L}((\phi_n, G_n, (Z_i)_{i=1}^{N_n})) \leq 1$ almost-surely, since for every n, N_n is finite almost surely. We say a node j is *Flip-Bad in* G_n if

$$\mathcal{L}((\phi_n, G_n, (Z_i)_{i=1}^{N_n})) \leq \mathcal{L}((\phi_n, G_n, (\tilde{Z}_i^{(j)})_{i=1}^{N_n})),$$

where $\tilde{Z}_i^{(j)} = Z_i$ if $i \neq j$ and $\tilde{Z}_j^{(j)} = -Z_j$. Thus, we use the term Flip-Bad, where a node is 'bad' if on flipping its community label, the observed data becomes more likely. We use this definition to reason about the maximum-likelihood estimator. More formally, for each node $i \in [1, N_n]$, the Maximum-Likelihood (ML) estimate of the community label is denoted as $\hat{Z}_i$ and satisfies

$$\begin{aligned}(\hat{Z}_i)_{i \in [1, N_n]} &= \arg \max_{z \in \{-1, +1\}^{N_n}} \mathbb{P}[(Z_i)_{i \in [1, N_n]} = z | \phi_n, G_n] \\ &= \arg \max_{z \in \{-1, +1\}^{N_n}} \mathcal{L}((\phi_n, G_n, z)) \end{aligned} \tag{D.23}$$

In words, it is the optimal estimate for the community labels given the observed data ϕ_n and G_n, where optimality refers to the fact that it minimizes the probability that the vector is $(\hat{Z}_i)_{i \in [1, N_n]}$ is different from the ground truth

$(Z_i)_{i\in[1,N_n]}$ among all possible estimators. More formally, the ML estimator $(\hat{Z}_i)_{i\in[1,N_n]}$ in Equation (D.23) satisfies

$$\mathbb{P}[(\hat{Z}_i)_{i\in[1,N_n]} \neq (Z_i)_{i\in[1,N_n]}] = \inf_{(\tilde{Z}_i)_{i\in[1,N_n]}=\mathcal{A}(\phi_n,G_n)} \mathbb{P}[(\tilde{Z}_i)_{i\in[1,N_n]} \neq (Z_i)_{i\in[1,N_n]}],$$

where the infimum is over all measurable functions $\mathcal{A}$ of the data ϕ_n and G_n. This asserts that the ML estimator for the community labels is the optimal estimator that minimizes probability of error. The following proposition gives a structural condition on the model parameters of G_n to identify a sufficient condition when the ML estimator will fail.

Proposition 124. If the model parameters of G_n satisfies

$$\limsup_{n\to\infty} \frac{\int_{y\in B_n} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy)}{n\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n\cup\{0\}}]^2} \leq 1, \tag{D.24}$$

$$\lim_{n\to\infty} n\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n\cup\{0\}}] = \infty \tag{D.25}$$

where $m_{n,d}$ is the Haar measure on the torus B_n, then Exact-Recovery is not solvable.

The condition in Equation (D.24) states that the event of two 'far-away' nodes being Flip-Bad are asymptotically uncorrelated. Such statements are true for instance if the functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ have support that is $o(n)$. For example, if the connection functions satisfy $f_{in}^{(n)}(r) = a_n\mathbf{1}_{r\leq R_n}$ and $f_{out}^{(n)}(r) = b_n\mathbf{1}_{r\leq R_n}$, for some $0 \leq b_n \leq a_n \leq 1$ and $R_n = o(n^{1/d})$, then they satisfy the condition. See for example also Lemma 126. On the other hand, Equation (D.24) is satisfied even if the model is reduced to the classical

SBM without geometry, i.e., any two nodes in G_n are connected only based on their community labels and not the location labels if the average degree is proportional to the logarithm of the population size.

Proof. If there exists a node in G_n that is Flip-Bad, then the Maximum Likelihood estimator for the community labels will not match the ground truth. We will show that if Equations (D.24) and (D.25) hold, then $\mathbb{P}[\exists i \in [1, N_n] : i$ is Flip-Bad in $G_n] = 1 - o_n(1)$, which will conclude the proof of the result.

To do this, define by $Y_n := \sum_{i \in [1, N_n]} \mathbf{1}_{i \text{ is Flip-Bad in } G_n}$. From the classical method of moments, it suffices to establish that $\limsup_{n \to \infty} \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} \leq 1$ and $\mathbb{E}[Y_n] \to \infty$ as $n \to \infty$. Indeed, if this were the case, then by Chebychev's inequality, we would have

$$\mathbb{P}[Y_n = 0] \leq \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} - 1,$$

which will converge to 0. It remains to compute the first and second moment of Y_n.

From Campbell's theorem (see Appendix D.5 for Palm probability on the torus), we have

$$\begin{aligned}
\mathbb{E}[Y_n] &= \lambda \int_{y \in B_n} \mathbb{E}^y[\mathbf{1}_{y \text{ is Flip-Bad in } G_n \cup \{y\}}] m_{n,d}(dy) \\
&= \lambda n \mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-bad in } G_n \cup \{0\}}],
\end{aligned} \qquad (D.26)$$

where the second inequality follows from the symmetry in the torus. Thus from the hypothesis of the theorem in Equation (D.25), $\mathbb{E}[Y_n]$ converges to ∞.

283

To compute the second moment, we use the factorial moment expansion of the Poisson Process as follows.

$$\mathbb{E}[Y_n^2] \tag{D.27}$$

$$= \mathbb{E}\Big[\Big(\sum_{i\in[1,N_n]} \mathbf{1}_{i \text{ is Flip-Bad in } G_n}\Big)^2\Big]$$

$$= \mathbb{E}[Y_n] + \mathbb{E}\Big[\sum_{i\neq j} \mathbf{1}_{i \text{ is Flip-Bad in } G_n} \mathbf{1}_{j \text{ is Flip-Bad in } G_n}\Big]$$

$$\overset{(a)}{=} \mathbb{E}[Y_n] + \tag{D.28}$$

$$\lambda^2 \int_{x\in B_n} \int_{y\in B_n} \mathbb{E}^{x,y}\big[\mathbf{1}_{x \text{ is Flip-Bad in } G_n \cup \{x,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{x,y\}}\big] m_{n,d}(dx) m_{n,d}(dy)$$

$$\overset{(b)}{=} \mathbb{E}[Y_n] + \lambda^2 n \int_{y\in B_n} \mathbb{E}^{0,y}\big[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}\big] m_{n,d}(dy),$$

$$\tag{D.29}$$

where equality (a) follows from the 2nd order Moment Measure expansion of a Poisson process and (b) follows from the symmetry in the torus. Thus, from Equations (D.26) and (D.29) , we have,

$$\frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} = \frac{1}{\mathbb{E}[Y_n]} +$$

$$\frac{\lambda^2 \int_{y\in B_n} \mathbb{E}^{0,y}\big[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}\big] m_{n,d}(dy)}{\big(\lambda n \mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-bad in } G_n \cup \{0\}}]\big)^2}. \tag{D.30}$$

From the assumption on the connection functions in Equation (D.24) and the first moment in Equation (D.30), we have $\limsup_{n\to\infty} \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]} \leq 1$. □

In the rest of the section, we consider the model given in Definition 57 where the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ take the form $f_{in}(r) = a\mathbf{1}_{r\leq\log(n)^{1/d}}$ and $f_{out}^{(n)}(r) = b\mathbf{1}_{r\leq\log(n)^{1/d}}$ for some $0 \leq b < a \leq 1$, to illustrate

how one can use the previous proposition to obtain a closed form expression for the phase-transition threshold. We first prove Theorem 58 to provide an exact necessary condition in terms of the model parameters λ, a and b for achieving exact-recovery. The proof of Theorem 58 follows from the next two lemmas. Recall that, we denote by ν_d for all $d \in \mathbb{N}$ as the volume of the unit Euclidean ball in d dimensions.

Lemma 125. For all $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$, if $G_n \sim \mathcal{G}(\lambda n, a, b, d)$, then

$$\mathbb{E}^0\left[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}\right] = e^{-\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)} - o(1))}.$$

Lemma 126. For all $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, the graph $G_n \sim \mathcal{G}(\lambda n, a, b, d)$ satisfies Equation (D.24).

Proof. of Theorem 58.

From Lemma 126, we know that the connection functions satisfy the model assumption in Equation (D.24). From Lemma 125, we know that if $\lambda \nu_d(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, then $\mathbb{E}^0\left[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}\right] = n^{-1+\delta-o(1)}$ for some $\delta > 0$. Thus, $n\mathbb{E}^0\left[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}\right] = n^{\delta-o(1)}$ which converges to ∞ as n goes to ∞. The proof is now complete thanks to Proposition 124. $\square$

We now provide the proofs of the associated lemmas.

Proof. of Lemma 125.

This lemma is a corollary of Lemma 11 proven in [20], where the error expo-

285

nent for hypothesis testing between Poisson random vectors was established. To assess whether the node at 0 is Flip-Bad, we need to decide given the locations and true community labels of the neighbors and non-neighbors of node at 0, whether it belongs to community $+1$ or -1. However, since the connection functions have support of $\log(n)^{1/d}$ and events in disjoint regions of space are independent, it suffices to consider the neighbors and non-neighbors within the ball of radius $\log(n)^{1/d}$ around 0. (Note from the model that there are no neighbors of the node at 0 at a distance larger than $\log(n)^{1/d}$). The number of neighbors of nodes in the same commuity as 0 is a Poisson random variable of mean $\lambda/2\nu_d a \log(n)$ and the number of neighbors in the opposite community is another independent Poisson random variable of mean $\lambda/2\nu_d b \log(n)$. The independence follows from elementary independent thinning property of the PPP. Similarly, the number of non-neighbors in the same and opposite community as 0 and within a distance of $\log(n)^{1/d}$ of 0 are independent Poisson random variables of mean $\lambda/2\nu_d(1-a)\log(n)$ and $\lambda/2\nu_d(1-b)\log(n)$ respectively. This argument again follows from the independent thinning property of an independently marked Poisson Process.

Thus the probability of a node at 0 being flip-bad is equal to the error made by an optimal hypothesis tester between two random vectors $(\lambda/2)\nu_d \log(n)(a, b, 1-a, 1-b)$ and $(\lambda/2)\nu_d \log(n)(b, a, 1-b, 1-a)$, given an uniform prior over the two models. Furthermore, the components of the observed random vector are independent. Thus, we are in a setting to apply the CH-divergence theorem of [20] to characterize the error probability

286

of the optimal hypothesis testing error. More precisely, applying Lemma 11 from [20] the error probability in identifying between the two Poisson random vectors with uniform prior $(\lambda/2)\nu_d \log(n)\mu$ and $(\lambda/2)\nu_d \log(n)\nu$ where the vectors $\mu := (a, b, 1 - a, 1 - b)$ and $\nu := (b, a, 1 - b, 1 - a)$ is equal to $\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}]$ is given by

$$\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}] := n^{-(\lambda/2)\nu_d D_+(\mu,\nu) + o(1)}, \tag{D.31}$$

where the CH-Divergence $D_+(\mu, \nu)$ ([20]) is given by

$$D_+(\mu, \nu) := \max_{t \in [0,1]} \sum_{x \in \mathcal{X}} (t\mu(x) + (1 - t)\nu(x) - \mu(x)^t \nu(x)^{1-t}). \tag{D.32}$$

Here $\mathcal{X} := [1, 2, 3, 4]$, and $\mu(x)$ $(\nu(x))$ for $x \in \mathcal{X}$ refers to the xth component of the vector μ (ν). Evaluating Equation (D.32) yields that $D_+(\mu, \nu) = 2(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)})$ with the maximum being achieved at $t = 1/2$ due to symmetry in the vectors μ and ν. Substituting Equation (D.32) into (D.31) yields the result. □

Proof. of Lemma 126.

This lemma follows from some straightforward calculations exploiting the spatial independence across the Poisson process. To verify Equation

(D.24), consider the following chain of equations.

$$\int_{y \in B_n} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy) =$$

$$\int_{y \in B(0, 2\log(n)^{1/d})} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy)$$

$$+ \int_{y \in B_n \cap B(0, 2\log(n)^{1/d})^{\complement}} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy)$$

$$\leq \int_{y \in B(0, 2\log(n)^{1/d})} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}}] m_{n,d}(dy)$$

$$+ \int_{y \in B_n \cap B(0, 2\log(n)^{1/d})^{\complement}} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy)$$

(D.33)

The key observation to make is that for x, y such that $||x - y|| > 2\log(n)^{1/d}$, we have

$$\mathbb{E}^{x,y} [\mathbf{1}_{x \text{ is Flip-Bad in } G_n \cup \{x,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{x,y\}}] =$$

$$\mathbb{E}^{x} [\mathbf{1}_{x \text{ is Flip-Bad in } G_n \cup \{x\}}] \mathbb{E}^{y} [\mathbf{1}_{y \text{ is Flip-Bad in } G_n \cup \{y\}}] \quad \text{(D.34)}$$

This follows since in an independently marked Poisson Process, events on disjoint sets are independent. Further more, from the symmetry in the torus, for all $x \in B_n$, we also have

$$\mathbb{E}^{x} [\mathbf{1}_{x \text{ is Flip-Bad in } G_n \cup \{x\}}] = \mathbb{E}^{0} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}]. \quad \text{(D.35)}$$

Thus, we get from Equations (D.33),(D.34) and (D.35) that

$$\int_{y \in B_n} \mathbb{E}^{0,y} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy) \leq$$

$$(n - 2^d \nu_d \log(n)) \mathbb{E}^{0} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}]^2 + 2^d \nu_d \log(n) \mathbb{E}^{0} [\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}].$$

(D.36)

288

Thus, since $\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, we have from Lemma 125 that,

$n\mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}] = n^\gamma$ for some $\gamma > 0$. Hence Equation (D.36) implies Equation (D.24) if $\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$.

$\square$

D.3.2   Upper Bound for Exact-Recovery - Proof of Theorem 60

In the present paper, we are only able to establish the presence of the phase transition by proving Theorem 60. We believe a 'two-round' information theoretic argument can be employed to prove this result. A possible strategy is to first show that for any $0 \leq b < a \leq 1$, a 'large' (i.e. all but $o(n)$) nodes will be correctly classified by the ML estimator with high probability. Further, if the parameters satisfied $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, then all nodes will be correctly classified with high probability. One can possibly make this efficient by means of 'sample splitting' arguments of [20]. One can sub-sample the edges so that the graph is almost sparse so that a large fraction of the nodes can be correctly labeled by the ML estimator. Then, we can 'clean-up', i.e. estimate a corrected community label estimate using the edges not used in the first round. This conjecture is also reminiscent of the 'local to global' phenomena that occurs in many random graph models ([18],[286],[83],[305]), where an obvious local necessary condition also turns out to be sufficient. However, as a corollary to the GBG algorithm introduced above, we Theorem 60 which establishes that Exact-Recovery can be solved if the intensity λ is

sufficiently high.

Proof. of Theorem 60

Notice that with $R = \log(n)^{1/d}/2d$, there are at-most $\lceil 4^d dn/\log(n) \rceil$ grid-cells in B_n. If we show that the chance that a grid cell is T-BAD is $n^{-1-\delta}$ for some $\delta > 0$, then by an union bound argument, we can assert that with probability at-least $1 - n^{-\delta}$, all grid-cells will be T-GOOD. Furthermore, if all grid cells are T-GOOD, then by Proposition 116, all nodes in G_n will be correctly partitioned by the algorithm.

Set $R = \log(n)^{1/d}/2d$ and $\epsilon \in (0,1)$ arbitrary as parameters of the GBG algorithm. From Proposition 118, we know that for any $z, z' \in \mathbb{Z}^d$ such that $||z - z'||_\infty = 1$, the probability that the pairwise classifier makes an error is at-most $e^{-c(d,a,b)\lambda \log(n)}$, where $c(d,a,b) > 0$ is a positive constant. From Lemma 119, it is clear, that the probability that a cell is T-BAD is at-most $n^{-c(d,a,b)\lambda} \log(n)^d C + n^{-c'(\epsilon)\lambda n}$, where c and c' are two strictly positive constants. Thus, by choosing λ sufficiently high, we can ensure that the probability of a grid-cell being T-BAD is at-most $n^{-1-\delta}$ for some $\delta > 0$. $\qquad\square$

Here we also want to remark that using ideas from Section 5.4.5, the GBG algorithm can be implemented without knowledge of the parameters. In particular, one can use standard spectral or SDP algorithms to decide whether a cell is T-GOOD or not, as explained in Section 5.4.5. This can be done so since the sub-graph of G_n restricted to nodes within the 1 thickening of any grid cell z is dsitributed as the SBM with random $\log(n)$ number of nodes and

having a constant connection probability of either a or b among nodes of the same or opposite communities. Thus, one can employ standard spectral or SDP algorithms 'off the shelf' to implement the algorithm without knowledge of a,b or λ.

D.4   Proof of Lemma 79

Proof. Denote by $T_I(j)$ be the breadth first spanning tree of I constructed with j as the root. Thus in the tree $T_I(j)$, and for all $k \in V_I(j)$, there is exactly one path from j to k in the tree $T_I(j)$.

The existence of two labelings is not so difficult since we know there exists one underlying true labeling which generated the data G, I. But since, the model is symmetric, the complement of the true labels will also be consistent in the sense of Lemma 78. Thus, there are at least two labeling consistent with the observed data G, I. These two labels of $V_I(j)$ can be constructed explicitly which we do in the next paragraph. We then show, that there are no other that can be consistent in the sense of Lemma 78, which will conclude the proof.

To construct the two possible labelings, first assume that $Z_j = +1$. Now conditionally on this and G, each neighbor of j in $T_I(j)$ will have exactly one possible community label estimate that is consistent in the sense of Lemma 78. Now, by induction, we can construct the labels of $V_I(j)$. Assume, that conditionally on $Z_j = +1$ and G, we have a unique set of labels for all vertices in $T_I(j)$ at graph distance of less than or equal to k. Let $u \in T_I(j)$ be an

arbitrary vertex such that it is at graph distance $k+1$ from j in $T_I(j)$. Since $T_I(j)$ is a tree, there is a unique vertex v in $V_I(j)$ such that $v \sim_{T_I(j)} u$ and v is at a distance of k from j. Thus, conditionally on $Z_j = +1$, Z_v is a fixed community label due to the induction hypothesis. Since Z_v is fixed, then there is a unique label for Z_u that will be consistent in the sense of Lemma 78. Since u was arbitrary, we can uniquely assign a community label to all vertices at graph distance of $k+1$ from j in $T_I(j)$. Hence, by induction, conditionally on $Z_j = +1$, there is a unique community estimate for all vertices in $V_I(j)$. Similarly, if we assumed $Z_j = -1$, we will find another unique labeling for the vertices in $V_i(j)$ which will be the complement of the unique labeling obtained by assuming $Z_j = +1$. This, gives us that there exist at-least two labelings of $V_i(j)$ that are complements of each other and consistent with the observed data G and I in the sense of Lemma 78.

To see that there can be no other possibilities, we argue by contradiction. Assume there are two labelings and a vertex k such that in one of the labelings $Z_j = +1, Z_k = +1$ and in the other $Z_j = +1, Z_k = -1$. It is clear that at-most one of the above labelings will be consistent in the tree $T_I(j)$ in the sense of Lemma 78. This establishes that the two sequences we constructed in the previous paragraph which are complements of each other are the only two possible sequences that are consistent in the sense of Lemma 78.

$\square$

D.5 Palm Measure on the Torus

In this section, we recap the basic properties of the Poisson Point Process on the torus. As the torus is a locally compact unimodular topological group, the definition of Palm measures and the moment measure equations of PPP on a torus follows directly from the theory described in [228]. We reproduce the key results needed for our paper here. Fix a $n \in \mathbb{N}$ and consider the d dimensional torus on the set $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2} \right]^d$ equipped with the Haar measure denoted as $m_{n,d}(\cdot)$ which is invariant under translations on the torus. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which we have a stationary independently marked PPP on the torus B_n of intensity $\lambda > 0$ with marks in an arbitrary Polish space $\mathcal{K}$. Denote by the atoms of this point process by $\phi_n := \{X_1, \cdots, X_{N_n}\}$ enumerated in an arbitrary manner and the corresponding marks as $\{K_1, \cdots, K_{N_n}\}$. From the definition of the PPP, N_n is a Poisson random variable of mean λn independent of everything else and conditional on N_n, $(X_i)_{i=1}^{N_n}$ are i.i.d. random variables that are uniformly distributed in the set B_n. Conditionally on N_n and $(X_i)_{i=1}^{N_n}$, the sequence $(K_i)_{i=1}^{N_n}$ are i.i.d. In this framework, for any $k \in \mathbb{N}$ and $x_1, \cdots, x_k \in B_n$, the Palm measure $\mathbb{P}^{x_1, \cdots, x_k}$ corresponds to adding fictitious atoms at locations $\{x_1, \cdots, x_k\}$ and equipping them with independent marks having the same law as K_1. In other words, thanks to Slivnyak's theorem ([228]), the atoms of the PPP under the Palm measure $\mathbb{P}^{x_1, \cdots, x_k}$ is $\phi_n \cup \{x_1, \cdots x_k\}$ where ϕ_n is the law of the point process under $\mathbb{P}$, with the marks of these additional points having the same distribution as that of K_1 and independent of everything else. For this frame-

work, for any function $f(\cdot) : B_n \to \mathbb{R}_+$ such that $\int_{x \in B_n} f(x) m_{n,d}(dx) < \infty$ we have the following version of Campbell's theorem ([228]) -

$$\mathbb{E}[\sum_{x \in \phi_n} f(x)] = \lambda \int_{x \in B_n} \mathbb{E}^x[f(x)] m_{d,n}(dx). \tag{D.37}$$

More generally, we have the following k-th order moment measure expansions of the Poisson process. Let $f(\cdot) : B_n^k \to \mathbb{R}_+$ be such that

$$\int_{x_1, \cdots, x_k} f(x_1, \cdots, x_k) \prod_{i=1}^{k} m_{n,d}(dx_i) < \infty.$$

Then, we have

$$\mathbb{E}[\sum_{\substack{x_1, \cdots, x_k \in \phi_n \\ \neq}} f(x_1, \cdots, x_k)] = \lambda^k \int_{x_1, \cdots, x_k \in B_n} \mathbb{E}^{x_1, \cdots, x_k}[f(x_1, \cdots, x_k)] \prod_{i=1}^{k} m_{n,d}(dx_i).$$

$$\tag{D.38}$$

We use Equations (D.37) and (D.38) in Section D.3 to compute the first and second moments of the number of Flip-Bad nodes in G_n.

Bibliography

[1] Pirate Parties International. $http://www.pp\text{-}international.net/$.

[2] Planetlab. $http://www.planet\text{-}lab.org/$.

[3] Wake-on-LAN [accessed October 2012]. `http://en.wikipedia.org/wiki/Wake-on-LAN`.

[4] Wireless LAN medium access control (MAC) and physical layer (PHY) specification, 1997.

[5] Fiber channel physical interface (FC-PI), 1999.

[6] 3rd generation partnership project; technical specification group radio access network; physical layer aspects of utra high speed downlink packet access, March 2001.

[7] Results of l3nqs simulation study, October 2001.

[8] Cdma2000 high rate packet data air interface specification, August 2005.

[9] Medium access control (mac) standard for cdma2000 spread spectrum systems, September 2005.

[10] Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems–Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2005.

[11] Physical layer standard for cdma2000 spread spectrum systems, September 2005.

[12] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Feasibility study for evolved UTRA and UTRAN (Release 7), September 2006.

[13] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Physical layer aspects for evolved UTRA (Release 7), September 2006.

[14] Commission endorses, with comments, Spanish regulator's measure to make mobile market more competitive. *European Comission IP/06/97*, January 2006.

[15] Virtually mobile: What drives MVNO success. *McKinsey & Company*, June 2014.

[16] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[17] Emmanuel Abbe, Francois Baccelli, and Abishek Sankararaman. Community detection on euclidean random graphs. *arXiv preprint arXiv:1706.09942*, 2017.

[18] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[19] Emmanuel Abbe, Laurent Massoulie, Andrea Montanari, Allan Sly, and Nikhil Srivastava. Group synchronization on grids. *arXiv preprint arXiv:1706.08561*, 2017.

[20] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

[21] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[22] A. Abdrabou and W. Zhuang. Service time approximation in IEEE 802.11 single-hop ad hoc networks. 7(1):305–313, January 2008.

[23] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. *Self-Similar Network Traffic and Performance Evaluation,* Wiley, pages 39–88, 2000.

[24] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Self-similarity and long-range dependence through the wavelet lens. *Theory and Applications of Long-Range Dependence,* Birkhäuser Boston, pages 527–556, 2003.

[25] Reza Aghajani and Kavita Ramanan. The hydrodynamic limit of a randomized load balancing network. *arXiv preprint arXiv:1707.02005*,

2017.

[26] R. Agrawal and B. Subramanian. Optimality of certain channel aware scheduling policies. In *Proc. 40th Annual Allerton Conf. Comm., Control, and Comput.*, Monticello, IL, October 2002.

[27] Derek Aguiar and Sorin Istrail. Hapcompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. *Journal of Computational Biology*, 19(6):577–590, 2012.

[28] Soyeon Ahn and Haris Vikalo. Joint haplotype assembly and genotype calling via sequential monte carlo algorithm. *BMC bioinformatics*, 16(1):223, 2015.

[29] S. Albin. On Poisson approximations for superposition arrival processes in queues. *Mgmt. Sci.*, 28(2):126–137, 1982.

[30] S. L. Albin. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Oper. Res.*, 32(5):1133–1162, 1984.

[31] Venkat Anantharam and François Baccelli. The boolean model in the shannon regime: Three thresholds and related asymptotics. *arXiv preprint arXiv:1408.1338*, 2014.

[32] J. Andrews, F. Baccelli, and R. Ganti. A tractable approach to coverage and rate in cellular networks. 59(11):3122–3134, November 2011.

[33] Jeffrey G Andrews, François Baccelli, and Radha Krishna Ganti. A tractable approach to coverage and rate in cellular networks. *IEEE Transactions on Communications*, 59(11):3122–3134, 2011.

[34] G. B. Arfken and H. J. Weber. *Mathematical Methods For Physicists*. Academic Press, 4th edition, 1995.

[35] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[36] Ehsan Aryafar, Alireza Keshavarz-Haddad, Michael Wang, and Mung Chiang. RAT selection games in HetNets. In *Proc. IEEE Infocom*, 2013.

[37] R. Aumann. *War and Peace*. Nobel Prize Lecture, The Royal Swedish Academy of Sciences, Stockholm, Sweden, 2005.

[38] R. Aumann and J. Drèze. Cooperative games with coalition structures. *International Journal of Game Theory*, 3:217–237, 1974.

[39] R. Aumann and L. Shapley. *Values of Non-Atomic Games*. Princeton University Press, 1974.

[40] F Baccelli and P Brémaud. *Elements of queuing theory: Palm-martingale calculus and stochastic recurrence*. New York: Springer-Verlag, 1994.

[41] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences.* Springer-Verlag, 2nd edition, 2003.

[42] François Baccelli and Bartlomiej Blaszczyszyn. *Stochastic geometry and wireless networks: Theory*, volume 1. Now Publishers Inc, 2009.

[43] François Baccelli and Bartlomiej Blaszczyszyn. *Stochastic geometry and wireless networks: Theory*, volume 1. Now Publishers Inc, 2009.

[44] François Baccelli and Bartlomiej Blaszczyszyn. *Stochastic Geometry and Wireless Networks, Volume 1: Theory*, volume 3. Foundations and Trends in Networking, 2010.

[45] François Baccelli and Bartlomiej Blaszczyszyn. *Stochastic Geometry and Wireless Networks, Volume 2: Applications*, volume 4. Foundations and Trends in Networking, 2010.

[46] François Baccelli, Bartlomiej Blaszczyszyn, and Paul Mühlethaler. An aloha protocol for multihop mobile wireless networks. *IEEE Transactions on Information Theory*, 52(2):421–436, 2006.

[47] François Baccelli, Bartlomiej Blaszczyszyn, and Chaman Singh. Analysis of a proportionally fair and locally adaptive spatial aloha in poisson networks. In *Proc. IEEE Infocom*, pages 2544–2552, 2014.

[48] François Baccelli, Alexander Borovkov, and Jean Mairesse. Asymptotic results on infinite tandem queueing networks. *Probability Theory and Related Fields*, 118(3):365–405, 2000.

[49] François Baccelli and Pierre Brémaud. *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*, volume 26. Springer Science & Business Media, 2013.

[50] François Baccelli, Héctor A Chang-Lara, and Sergey Foss. Shape theorems for poisson hail on a bivariate ground. *Advances in Applied Probability*, 48(2):525–543, 2016.

[51] François Baccelli, Abbas El Gamal, and David NC Tse. Interference networks with point-to-point codes. *Information Theory, IEEE Transactions on*, 57(5):2582–2596, 2011.

[52] François Baccelli and Sergey Foss. Poisson hail on a hot ground. *Journal of Applied Probability*, 48(A):343–366, 2011.

[53] François Baccelli and Serguei Foss. On the saturation rule for the stability of queues. *Journal of Applied Probability*, 32(2):494–507, 1995.

[54] François Baccelli and Anastasios Giovanidis. A stochastic geometry framework for analyzing pairwise-cooperative cellular networks. 14(2):794–808, February 2015.

[55] François Baccelli, Fabien Mathieu, and Ilkka Norros. Mutual service processes in rd existence and ergodicity. 2015.

[56] François Baccelli, Fabien Mathieu, and Ilkka Norros. Mutual service processes in euclidean spaces: existence and ergodicity. *Queueing Systems*, 86(1-2):95–140, 2017.

[57] François Baccelli, Fabien Mathieu, Ilkka Norros, and Rémi Varloot. Can P2P networks be super-scalable? In *IEEE INFOCOM*, pages 1753–1761, 2013.

[58] François Baccelli, Fabien Mathieu, Ilkka Norros, and Rémi Varloot. Can p2p networks be super-scalable? In *INFOCOM, 2013 Proceedings IEEE*, pages 1753–1761. IEEE, 2013.

[59] Adrian Baddeley, Pablo Gregori, Jorge Mateu, Radu Stoica, and Dietrich Stoyan. *Case studies in spatial point process modeling.* Springer, 2006.

[60] Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[61] S. Baek and G. de Veciana. Spatial model for energy burden balancing and data fusion in sensor networks detecting bursty events. 53(10):3615–3628, October 2007.

[62] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

[63] Vikas Bansal and Vineet Bafna. Hapcut: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159, 2008.

[64] Vikas Bansal, Aaron L Halpern, Nelson Axelrod, and Vineet Bafna. An mcmc algorithm for haplotype assembly from whole-genome sequence data. *Genome research*, 18(8):1336–1346, 2008.

[65] S. Baset, J. Reich, J. Janak, P. Kasparek, V. Misra, D. Rubenstein, and H. Schulzrinne. How green is IP-telephony? In *Proc. ACM Sigcomm Workshop on Green Networking*, August 2010.

[66] F. M. Bass. A new product growth model for consumer durables. *Management Science*, 15:215–227, 1969.

[67] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons Inc., 2nd edition, 1993.

[68] M. Benaïm and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65(11-12):823–838, November 2008.

[69] Emily Berger, Deniz Yorukoglu, Jian Peng, and Bonnie Berger. Haptree: A novel bayesian framework for single individual polyplotyping using ngs data. *PLoS computational biology*, 10(3):e1003502, 2014.

[70] G. Berger-Sabbatel, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau. Fairness and its impact on delay in 802.11 networks. In *Proc. IEEE Globecom*, November 2004.

[71] C. Berrou and A. Glavieux. Near optimum error correcting coding and decoding: Turbo codes. 44:1261–1271, October 1996.

[72] D. Bersekas and R. Gallager. *Data Networks*. Englewood Cliffs, N.J.: Prentice-Hall, 2nd edition, 1992.

[73] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer. Network densification: The dominant theme for wireless evolution into 5G. 52(2):82–89, 2014.

[74] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. 18(3):535–547, March 2000.

[75] G. Bianchi, I. Tinnirello, and L. Scalia. Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations. 19(4):28–34, July 2005.

[76] P. Billingsley. *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition, 1999.

[77] N. Bingham and R. Doney. Asymptotic properties of supercritical branching processes I: The Galton-Watson process. *Advances in Applied Probability*, 6(4):711–731, 1974.

304

[78] N. Bingham, C. Goldie, and J. Teugels. *Regular Variation.* Cambridge University Press, 1989.

[79] Bartłomiej Błaszczyszyn, Christian Rau, and Volker Schmidt. Bounds for clump size characteristics in the boolean model. *Advances in Applied Probability*, 31(4):910–928, 1999.

[80] Bartłomiej Błaszczyszyn, D Yogeshwaran, et al. On comparison of clustering properties of point processes. *Advances in Applied Probability*, 46(1):1–20, 2014.

[81] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting gibbs state for the ising model on the bethe lattice. *Journal of Statistical Physics*, 79(1):473–482, 1995.

[82] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[83] Béla Bollobás and Oliver Riordan. *Percolation.* Cambridge University Press, 2006.

[84] T. Bonald, S. Borst, and A. Proutière. Inter-cell scheduling in wireless data networks. In *Proc. European Wireless*, Cyprus, Greece, April 2005.

[85] Thomas Bonald, Sem Borst, Nidhi Hegde, and Alexandre Proutière. Wireless data performance in multi-cell scenarios. *ACM SIGMETRICS Performance Evaluation Review*, 32(1):378–380, 2004.

[86] Thomas Bonald, Sem Borst, Nidhi Hegde, and Alexandre Proutiére. Wireless data performance in multi-cell scenarios. *ACM SIGMETRICS Performance Evaluation Review*, 32(1):378–380, 2004.

[87] Paola Bonizzoni, Riccardo Dondi, Gunnar W Klau, Yuri Pirola, Nadia Pisanti, and Simone Zaccaria. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. *Journal of Computational Biology*, 23(9):718–736, 2016.

[88] R. R. Boorstyn, A. Kershenbaum, B. Maglaris, and V. Sahin. Throughput analysis in multihop CSMA packet radio networks. 35(3):267–274, March 1987.

[89] C. Bordenave, D. McDonald, and A. Proutiere. Random multi-access algorithm: A mean field analysis. In *Proc. 43th Annual Allerton Conf. Comm., Control, and Comput.*, Monticello, IL, September 2005.

[90] C. Bordenave, D. McDonald, and A. Proutiere. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Networks and Heterogeneous Media*, 5(1):31–62, March 2010.

[91] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[92] Charles Bordenave, David McDonald, and Alexandre Proutiere. Asymptotic stability region of slotted aloha. *Information Theory, IEEE Transactions on*, 58(9):5841–5855, 2012.

[93] Alexandr A Borovkov, Dmitrii Korshunov, and Rolf Schassberger. Ergodicity of a polling network with an infinite number of stations. *Queueing Systems*, 32(1-3):169–193, 1999.

[94] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. 13(3):636–647, June 2005.

[95] Mic Bowman, Saumya K. Debray, and Larry L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[96] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambirdge University Press, 2004.

[97] Johannes Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[98] D. Bradley and R. Gupta. On the distribution of the sum of n non-identically distributed uniform random variables. *Ann. Inst. Statist. Math.*, 54(3):689–700, 2002.

[99] Maury Bramson, Yi Lu, and Balaji Prabhakar. Randomized load balancing with general service time distributions. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 275–286. ACM, 2010.

[100] M. Bredel and M. Fidler. Understanding fairness and its impact on quality of service in IEEE 802.11. In *Proc. IEEE Infocom*, April 2009.

[101] P. Brémaud. A Swiss army formula of Palm calculus. *Journal of Applied Probability*, 30(1):40–51, 1993.

[102] P. Brémaud. *Markov Chains, Gibbs Field, Monte Carlo Simulation and Queues*. Springer-Verlag, 1999.

[103] T. Bu, L. Li, and R. Ramjee. Generalized proportional fair scheduling in third generation wireless data networks. In *Proc. IEEE Infocom*, Barcelona, Spain, April 2006.

[104] Changxiao Cai, Sujay Sanghavi, and Haris Vikalo. Structured low-rank matrix factorization for haplotype assembly. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):647–657, 2016.

[105] M. Carvalho and J. Garcia-Luna-Aceves. A scalable model for channel access protocols in multihop ad hoc networks. In *Proc. ACM MobiCom*, September 2004.

[106] A. Casajus. On the stabiilty of coalition structures. *Economics Letters*, 100(2):271–274, 2008.

[107] A. Casajus. Outside options, componenet efficiency, and stability. *Games and Economic Behavior*, 65:49–61, 2009.

[108] A. Casajus and A. Tutic. *Nash bargaining, Shapley threats, and outside options*. Working paper, Chair of Microeconomics, University of Leipzig, Germany, 2008.

[109] A. Castaldini, A. Cavallini, B. Fraboni, P. Fernandez, and J. Piqueras. Midgap traps related to compensation processes in CdTe alloys. *Phys. Rev. B.*, 56(23):14897–14900, 1997.

[110] M. Cha, P. Rodriguez, S. Moon, and J. Crowcroft. On next-generation telco-managed P2P TV architectures. In *Proc. USENIX International Workshop on Peer-to-Peer Systems (IPTPS)*, February 2008.

[111] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: Spatial mean field regime. In *Proc. ACM Sigmetrics*, June 2009.

[112] Mark J Chaisson, Sudipto Mukherjee, Sreeram Kannan, and Evan E Eichler. Resolving multicopy duplications de novo using polyploid phasing. In *International Conference on Research in Computational Molecular Biology*, pages 117–133. Springer, 2017.

[113] D. Chase. Code combining—a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets. 33:385–393, May 1985.

[114] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[115] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. Quantifying Skype user satisfaction. In *Proc. ACM Sigcomm*, September 2006.

[116] Yuxin Chen, Govinda Kamath, Changho Suh, and David Tse. Community recovery in graphs with locality. In *International Conference on Machine Learning*, pages 689–698, 2016.

[117] Zhi-Zhong Chen, Fei Deng, and Lusheng Wang. Exact algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 29(16):1938–1945, 2013.

[118] F. Chiti, R. Fantacci, T. Pecorella, and F. Versaci. Soft combining Hybrid ARQ techniques application to 3G wireless packet networks. In *Proc. IEEE ICC*, pages 2139–2143, Anchorage, AK, May 2003.

[119] Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[120] J. Cho and Y. Jiang. Fundamentals of the backoff process in 802.11: Dichotomy of the aggregation. 61(4):1687–1701, April 2015. DOI: 10.1109/TIT.2015.2404795.

[121] J. Cho, J.-Y. Le Boudec, and Y. Jiang. On the asymptotic validity of the decoupling assumption for analyzing 802.11 MAC protocol. 58(11):6879–6893, November 2012. DOI: 10.1109/TIT.2012.2208582.

[122] J. Cho and Y. Yi. On the Shapley-like payoff mechanisms in peer-assisted services with multiple content providers. In *International Conference on Game Theory for Networks (GameNets), (An extended version is under 1st revision for publication in IEEE/ACM Trans. Networking. available at* `http: // arxiv. org/ abs/ 1104. 0458` *)*, April 2011.

[123] J. Cho and Y. Yi. On the payoff mechanisms in peer-assisted services with multiple content providers: Rationality and fairness. 22(3):731–744, June 2014. DOI: 10.1109/TNET.2013.2259637.

[124] Cisco Systems, Incorporated. Cisco visual networking index: Forecast and methodology, 2010-2015. *White Paper, available at* `http: // www. cisco. com/ en/ US/ netsol/ ns827/ networking_ solutions_ white_ papers_ list. html`, June 2011.

[125] Andrew G Clark. The role of haplotypes in candidate gene studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 27(4):321–333, 2004.

[126] Malcolm Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[127] Jacob Cohen and Onno Boxma. *Boundary value problems in queueing system analysis*, volume 79. Elsevier, 2000.

[128] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[129] Potato Genome Sequencing Consortium et al. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189, 2011.

[130] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[131] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[132] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. 5(6):835–846, December 1997.

[133] Mark E Crovella and Azer Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[134] B. D. Cullity. *Introduction to Magnetic Materials.* Addison-Wesley, Reading, MA, 1972.

[135] Jim G Dai. On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, pages 49–77, 1995.

[136] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure.* Springer Science & Business Media, 2007.

[137] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes - vol. I: Elementary Theory and Methods.* Springer-Verlag, 2nd edition, 2003.

[138] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes - vol. II: General Theory and Structure.* Springer-Verlag, 2nd edition, 2008.

[139] Aleksandar Damnjanovic, Juan Montojo, Yongbin Wei, Tingfang Ji, Tao Luo, Madhavan Vajapeyam, Taesang Yoo, Osok Song, and Durga Malladi. A survey on 3gpp heterogeneous networks. *Wireless Communications, IEEE*, 18(3):10–21, 2011.

[140] Shreepriya Das and Haris Vikalo. Sdhap: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC genomics*, 16(1):260, 2015.

[141] A. De Meyer. On a theorem of Bingham and Doney. *J. Appl. Probab.*, 19(1):217–220, 1982.

[142] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for mod-

ular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[143] Fei Deng, Wenjuan Cui, and Lusheng Wang. A highly accurate heuristic algorithm for the haplotype assembly problem. *BMC genomics*, 14(2):S2, 2013.

[144] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491, 2011.

[145] Xavier Descombes. *Stochastic geometry for image analysis.* John Wiley & Sons, 2013.

[146] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews. Modeling and analysis of K-tier downlink heterogeneous cellular networks. 30(3):550–560, April 2012.

[147] RL Dobrushin and SB Shlosman. Completely analytical interactions: constructive description. *Journal of Statistical Physics*, 46(5-6):983–1014, 1987.

[148] Jorge Duitama, Thomas Huebsch, Gayle McEwen, Eun-Kyung Suk, and Margret R Hoehe. Refhap: a reliable and fast algorithm for single individual haplotyping. In *Proceedings of the First ACM International*

Conference on Bioinformatics and Computational Biology, pages 160–169. ACM, 2010.

[149] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6):469–493, 1992.

[150] Richard Durrett. *Lecture notes on particle systems and percolation.* Brooks/Cole Pub Co, 1988.

[151] M. Durvy, O. Dousse, and P. Thiran. Modeling the 802.11 protocol under different capture and sensing capabilities. In *Proc. IEEE Infocom*, Anchorage, AK, May 2007.

[152] M. Durvy, O. Dousse, and P. Thiran. Border effects, fairness, and phase transition in large wireless networks. In *Proc. IEEE Infocom*, Pheonix, AZ, April 2008.

[153] M. Durvy and P. Thiran. A packing approach to compare slotted and non-slotted medium access control. In *Proc. IEEE Infocom*, Barcelona, Spain, April 2006.

[154] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world.* Cambridge University Press, 2010.

[155] Abbas El Gamal and Young-Han Kim. *Network information theory.* Cambridge university press, 2011.

[156] William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

[157] Guy Fayolle and Roudolf Iasnogorodski. Two coupled processors: the reduction to a riemann-hilbert problem. *Probability Theory and Related Fields*, 47(3):325–351, 1979.

[158] W. Feller. *An Introduction to Probability and Its Applications, vol. 1*. Wiley, New York, 3rd edition, 1968.

[159] W. Feller. *An Introduction to Probability and Its Applications, vol. 2*. Wiley, New York, 2nd edition, 1971.

[160] PA Ferrari and Luiz Renato G Fontes. The net output process of a system with infinitely many queues. *The Annals of Applied Probability*, pages 1129–1144, 1994.

[161] D. Figueiredo, B. Liu, A. Feldmann, V. Misra, D. Towsley, and W. Willinger. On TCP and self-similar traffic. *Perf. Eval.*, 61(2-3):129–141, July 2005.

[162] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[163] Sergei Foss and Denis Denisov. On transience conditions for markov chains. *Siberian Mathematical Journal*, 42(2):364–371, 2001.

[164] Sergey Foss, Takis Konstantopoulos, and Thomas Mountford. Power law condition for stability of poisson hail. *Journal of Theoretical Probability*, pages 1–21, 2014.

[165] S.G. Foss and N.I. Chernova. On polling systems with infinitely many stations. *Siberian Mathematical Journal*, 37(4):832–846, 1996.

[166] F. Frederiksen and T. E. Kolding. Performance and modeling of WCDMA/HSDPA transmission/H-ARQ schemes. In *Proc. IEEE VTC Fall*, pages 472–476, Vancouver, Canada, September 2002.

[167] P. Frenger, S. Parkvall, and E. Dahlman. Performance comparison of HARQ with Chase combining and incremental redundancy for HSDPA. In *Proc. IEEE VTC Fall*, pages 1829–1833, Atlantic City, NJ, September 2001.

[168] N. M. Freris, H. Kowshik, and P. R. Kumar. Fundamentals of large sensor networks: Connectivity, capacity, clocks, and computation. 98(11):1828–1846, 2010.

[169] R. Gaigalas. A Poisson bridge between fractional Brownian motion and Lévy motion. *Stochastic Processes and their Applications*, 116:447–462, 2006.

[170] R. Gaigalas and I. Kaj. Convergence of scaled renewal processes and a packet arrival model. *Bernoulli*, 9(4):671–703, 2003.

[171] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. *arXiv preprint arXiv:1709.05510*, 2017.

[172] Nancy L Garcia and Thomas G Kurtz. Spatial birth and death processes as solutions of stochastic equations. *Alea*, 1:281–303, 2006.

[173] M. Garetto, T. Salonidis, and E. Knightly. Modeling per-flow through-put and capturing starvation in CSMA multi-hop wireless networks. 16(4):864–877, August 2008.

[174] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

[175] N. Gast, B. Gaujal, and J.-Y. Le Boudec. Mean field for Markov decision processes: from discrete to continuous optimization. 57(8), 2012.

[176] Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter, 2011.

[177] Filippo Geraci. A comparison of several algorithms for the single individual snp haplotyping reconstruction problem. *Bioinformatics*, 26(18):2217–2225, 2010.

[178] A. Gjendemsjø, D. Gesbert, G. E. Øien, and S. G. Kiani. Optimal power allocation and scheduling for two-cell capacity maximization. In *Proc. RAWNET*, Boston, MA, April 2006.

[179] Andrea Goldsmith. *Wireless Communications.* Cambridge University Press, 2005.

[180] Carl Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37(1):198–211, 2000.

[181] H Christian Gromoll. Diffusion approximation for a processor sharing queue in heavy traffic. *The Annals of Applied Probability*, pages 555–611, 2004.

[182] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[183] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of Bittorrent-like systems. In *Proc. ACM/USENIX IMC*, October 2005.

[184] Piyush Gupta and Panganmala R Kumar. The capacity of wireless networks. *IEEE Transactions on information theory*, 46(2):388–404, 2000.

[185] R. K. Gupta and S. D. Senturia. Pull-in time dynamics as a measure of absolute pressure. In *Proc. IEEE International Workshop on Microelectromechanical Systems (MEMS'97)*, pages 290–294, Nagoya, Japan, January 1997.

[186] M. Haenggi, J. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. 27(7):1029–1044, September 2009.

[187] Martin Haenggi, Jeffrey G Andrews, François Baccelli, Olivier Dousse, and Massimo Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1029–1046, 2009.

[188] J. Hagenauer. Rate-compatible punctured convolutional codes (RCPC codes) and their applications. 36:389–400, April 1988.

[189] Bruce Hajek. Balanced loads in infinite networks. *The Annals of Applied Probability*, 6(1):48–75, 1996.

[190] Khairi Ashour Hamdi. A useful lemma for capacity analysis of fading interference channels. *IEEE Transactions on Communications*, 58(2):411–416, 2010.

[191] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[192] S. Hart and M. Kurz. Endogenous formation of coalitions. *Econometrica*, 51:1047–1064, 1983.

[193] Abolfazl Hashemi, Banghua Zhu, and Haris Vikalo. Sparse tensor decomposition for haplotype assembly of diploids and polyploids. *BMC genomics*, 19(4):191, 2018.

[194] Dan He, Arthur Choi, Knot Pipatsrisawat, Adnan Darwiche, and Eleazar Eskin. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 26(12):i183–i190, 2010.

[195] J. He, A. Chaintreau, and C. Diot. A performance evaluation of scalable live video streaming with nano data centers. *Computer Networks*, 53:153–167, 2009.

[196] Maurice Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[197] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[198] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[199] J. Horrigan. Home broadband adoption 2008. *Pew Internet & American Life Project*, July 2008.

[200] L. Hu, M. Vojnović, and J.-Y. Le Boudec. Optimal channel choice for collaborative ad-hoc dissemination. In *Proc. IEEE Infocom*, March 2010.

[201] J. Jeong, Y. Yi, J. Cho, D. Eun, and S. Chong. Energy-efficient Wi-Fi sensing policy under generalized mobility patterns with aging. *under revision for IEEE/ACM Trans. Networking*, December 2014. (an eariler version in *IEEE Infocom* 2013).

[202] T. Ji and W. Stark. Rate-adaptive transmission over correlated fading channels. 53:1663–1670, October 2005.

[203] L. Jiang and J. Walrand. A distributed CSMA algorithm for throughput and utility maximization in wireless networks. accepted for publication.

[204] I. Kaj. Limiting fractal random processes in heavy-tailed systems. *Fractals in Engineering,* Springer-Verlag London, pages 199–218, 2005.

[205] I. Kaj and M. Taqqu. Convergence to fractional Brownian motion and to the Telecom process: the integral representation appoach. *Progress in Probability,* Birkhäuser Verlag Basel, 60:383–427, 2008.

[206] M. Kemal Karakayali, Gerard J. Foschini, and Reinaldo A. Valenzuela. Network coordination for spectrally efficient communications in cellular systems. 13:56–61, August 2006.

[207] Bruno Kauffmann, Francois Baccelli, Augustin Chaintreau, Vivek Mhatre, Konstantina Papagiannaki, and Christophe Diot. Measurement-

based self organization of interfering 802.11 wireless access networks. In *Proc. IEEE Infocom*, Anchorage, AK, May 2007.

[208] F. Kelly, A. Maullo, and D. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, July 1998.

[209] F. P. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, New York, NY, 2011.

[210] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[211] M Ya Kel'bert, ML Kontsevich, and AN Rybko. On Jackson networks on denumerable graphs. *Theory of Probability & Its Applications*, 33(2):358–361, 1989.

[212] M. Khouzani, S. Sarkar, and E. Altman. Maximum damage malware attack in mobile wireless networks. In *Proc. IEEE International Conference on Computer Communications (Infocom)*, March 2010.

[213] Saad G. Kiani and David Gesbert. Optimal and distributed scheduling for multicell capacity maximization. December 2006.

[214] Jong Hyun Kim, Michael S Waterman, and Lei M Li. Diploid genome reconstruction of ciona intestinalis and comparative analysis with ciona savignyi. *Genome research*, 17(7):000–000, 2007.

[215] Y. Kim, J. k. Han, H. Kwon, and D. Kim. Performance of HARQ and the effect of imperfect power control in CDMA reverse link. In *Proc. IEEE Globecom*, pages 3017–3021, Dallas, TX, November 2004.

[216] R. Knopp and P. A. Humblet. Information capacity and power control in single cell multiuser communications. In *Proc. IEEE ICC*, pages 331–335, Seattle, WA, June 1995.

[217] C. Koksal, H. Kassab, and H. Balakrishnan. An analysis of short-term fairness in wireless media access protocols. In *Proc. ACM Sigmetrics*, June 2000.

[218] T. E. Kolding, F. Frederiksen, and P. E. Mogensen. Performance aspects of WCDMA systems with high speed downlink packet access (HSDPA). In *Proc. IEEE VTC Fall*, pages 477–481, Vancouver, Canada, September 2002.

[219] Volodymyr Kuleshov. Probabilistic single-individual haplotyping. *Bioinformatics*, 30(17):i379–i385, 2014.

[220] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs. 15(3):588–601, June 2007.

[221] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer networks*, 31(11):1481–1493, 1999.

[222] K. Kumaran and L. Qian. Uplink scheduling in CDMA packet-data systems. In *Proc. IEEE Infocom*, pages 292–300, San Francisco, CA, March 2003.

[223] H. J. Kushner and P. A. Whiting. Convergence of proportional-fair sharing algorithms under general conditions. 3:1250–1259, July 2004.

[224] B.-J. Kwak, N.-O. Song, and L. Miller. Performance analysis of exponenetial backoff. 13(2):343–355, April 2005.

[225] Leslie Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[226] Giuseppe Lancia, Vineet Bafna, Sorin Istrail, Ross Lippert, and Russell Schwartz. Snps problems, complexity, and algorithms. In *European symposium on algorithms*, pages 182–193. Springer, 2001.

[227] Nate Lanxon. iPlayer uncovered: What powers the BBC's epic creation? *CNET UK*, May 2009.

[228] Günter Last. *Modern Random Measures: Palm Theory and Related Models*. Univ. Karlsruhe, Fak. für Mathematik, 2008.

[229] C. Lautensack and S. Zuyev. Random Laguerre tessellations. *Advances in Applied Probability*, 40(3):630–650, 2008.

[230] J.-Y. Le Boudec. Performance evaluation lecture notes. *available at http://perfeval.epfl.ch*, September 2009.

325

[231] J.-Y. Le Boudec. *Performance Evaluation of Computer and Communication Systems.* EPFL Press, 2010.

[232] J.-Y. Le Boudec and M. Vojnović. The random trip model: Stability, stationary regime, and perfect simulation. 14(6):1153–1166, December 2006.

[233] J.-Y. LeBoudec, D. McDonald, and J. Mundinger. A generic mean field convergence result for systems of interacting objects. In *Proc. 4th International Conference on the Quantitative Evaluation of SysTems (QEST)*, Edinburgh, UK, September 2007.

[234] H. Lee, H. Jang, J. Cho, and Y. Yi. On the stability of ISPs' coalition structure: Shapley value based revenue sharing. In *Proc. Conference on Information Sciences and Systems (CISS)*, available at `http://dx.doi.org/10.1109/CISS.2012.6310933`, Princeton University, New Jersey, March 2012.

[235] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. In *Proc. ACM Sigcomm*, September 1993.

[236] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection.

[237] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gen-

nady Denisov, et al. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, 2007.

[238] Guoqin Li and Hui Liu. Downlink radio resource allocation for multi-cell OFDMA system. 5:3451–3459, December 2006.

[239] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[240] Y. Li, F. Baccelli, H. Dhillon, and J. Andrews. Statistical modeling and probabilistic analysis of cellular networks with determinantal point processes. *submitted to IEEE Trans. Commun.*, December 2014. (available at http://arxiv.org/abs/1412.2087).

[241] Thomas Liggett. *Interacting particle systems*, volume 276. Springer Science & Business Media, 2012.

[242] Thomas M Liggett, Roberto H Schonmann, and Alan M Stacey. Domination by product measures. *The Annals of Probability*, 25(1):71–95, 1997.

[243] Xiaojun Lin, Ness B Shroff, and R Srikant. On the connection-level stability of congestion-controlled communication networks. *IEEE Transactions on Information Theory*, 54(5):2317–2338, 2008.

[244] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[245] Ross Lippert, Russell Schwartz, Giuseppe Lancia, and Sorin Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in bioinformatics*, 3(1):23–31, 2002.

[246] L. Lipsky and P. Fiorini. Auto-correlation of counting processes associated with renewal processes. *Technical Report, Booth Research Center, University of Connecticut*, September 1995.

[247] B. Liu and D. Towsley. A study on the coverage of large-scale sensor networks. In *Proc. IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, October 2004.

[248] Hui Liu and Guoqing Li. *OFDM-Based Broadband Wireless Networks: Design and Optimization*. Wiley-Interscience, 2005.

[249] Jiaping Liu, Alexandre Proutière, Yung Yi, Mung Chiang, and H Vincent Poor. Flow-level stability of data networks with non-convex and time-varying rate regions. In *ACM SIGMETRICS Performance Evaluation Review*, volume 35, pages 239–250, 2007.

[250] X. Liu, E. K. P. Chong, and N. B. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. 19(10):2053–2064, October 2001.

[251] Robert M Loynes. The stability of a queue with non-independent inter-arrival and service times. In *Mathematical Proceedings of the Cambridge*

Philosophical Society, volume 58, pages 497–520. Cambridge University Press, 1962.

[252] Eyal Lubetzky and Allan Sly. An exposition to information percolation for the ising model. *arXiv preprint arXiv:1501.00128*, 2014.

[253] Eyal Lubetzky and Allan Sly. Universality of cutoff for the ising model. *arXiv preprint arXiv:1407.1761*, 2014.

[254] Eyal Lubetzky and Allan Sly. Information percolation and cutoff for the stochastic ising model. *Journal of the American Mathematical Society*, 29(3):729–774, 2016.

[255] Russell Lyons and Yuval Peres. *Probability on trees and networks*, volume 42. Cambridge University Press, 2016.

[256] R. Ma, D. Chiu, J. Lui, V. Misra, and D. Rubenstein. On cooperative settlement between content, transit and eyeball Internet service providers. In *Proc. ACM CoNEXT*, December 2008.

[257] Jean Mairesse and Balaji Prabhakar. The existence of fixed points for the ·/GI/1 queue. *The Annals of Probability*, pages 2216–2236, 2003.

[258] VA Malyshev and IP Tsaregradskii. A queueing system with local interaction. *Theory of Probability & Its Applications*, 27(3):619–624, 1983.

[259] James B Martin. Large tandem queueing networks with blocking. *Queueing Systems*, 41(1-2):45–72, 2002.

[260] L. Massoulié and M. Vojnović. Coupon replication systems. 16(3):603–616, 2008.

[261] Laurent Massoulié. Structural properties of proportional fairness: stability and insensitivity. *The Annals of Applied Probability*, pages 809–839, 2007.

[262] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

[263] Laurent Massoulié and James Roberts. Bandwidth sharing: objectives and algorithms. In *IEEE INFOCOM*, volume 3, pages 1395–1403, 1999.

[264] Laurent Massoulié and James Roberts. Bandwidth sharing: objectives and algorithms. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1395–1403. IEEE, 1999.

[265] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[266] K. Medepalli and F. Tobagi. Towards performance modeling of IEEE 802.11 based wireless networks: A unified framework and its applications. In *Proc. IEEE Infocom*, April 2006.

[267] Ronald Meester, Mathew D Penrose, and Anish Sarkar. The random connection model in high dimensions. *Statistics & probability letters*, 35(2):145–153, 1997.

[268] Ronald Meester and Rahul Roy. *Continuum percolation*, volume 119. Cambridge University Press, 1996.

[269] D. Menasché, A. Rocha, E. Silva, D. Towsley, and R. Leão. Implications of peer selection strategies by publishers on the performance of P2P swarming systems. In *Proc. ACM Sigmetrics Workshop on MAthematical Performance Modeling and Analysis (MAMA)*, June 2011.

[270] D. Merugu, B. Prabhakar, and N. Rama. An incentive mechanism for decongesting the roads: A pilot program in Bangalore. In *Proc. ACM Workshop on the Economics of Networked Systems (NetEcon)*, July 2009.

[271] Sean P Meyn. Transience of multiclass queueing networks via fluid limit models. *The Annals of Applied Probability*, 5(4):946–957, 1995.

[272] V. Misra, S. Ioannidis, A. Chaintreau, and L. Massoulié. Incentivizing peer-assisted services: A fluid Shapley value approach. In *Proc. ACM Sigmetrics*, June 2010.

[273] M. Miyazawa. The derivation of invariance relations in complex queueing systems with stationary inputs. *Advances in Applied Probability*, 15(4):874–885, 1983.

[274] N. Miyoshi and T. Shirai. A cellular network model with Ginibre configured base stations. *Advances in Applied Probability*, 46(3):832–845, 2014.

[275] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. 8(5):556–567, October 2000.

[276] Ciamac Moallemi and Devavrat Shah. On the flow-level dynamics of a packet-switched network. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 83–94, 2010.

[277] Ciamac Moallemi and Devavrat Shah. On the flow-level dynamics of a packet-switched network. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 83–94. ACM, 2010.

[278] J. Møller. Random Johnson-Mehl tessellations. *Advances in Applied Probability*, 24(4):814–844, 1992.

[279] Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

[280] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. *arXiv preprint arXiv:1504.05910*, 2015.

[281] Cristopher Moore. The computer science and physics of community detection: landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

[282] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

[283] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[284] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[285] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[286] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75. ACM, 2015.

[287] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh An-*

nual ACM on Symposium on Theory of Computing, pages 69–75. ACM, 2015.

[288] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

[289] Ehsan Motazedi, Dick de Ridder, Richard Finkers, Samantha Baldwin, Susan Thomson, Katrina Monaghan, and Chris Maliepaard. Tripoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*, 2018.

[290] Ehsan Motazedi, Richard Finkers, Chris Maliepaard, and Dick de Ridder. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in bioinformatics*, 19(3):387–403, 2017.

[291] R. Myerson. Graphs and cooperation in games. *Mathematics of Operations Research*, 2:225–229, 1977.

[292] S. Nedevschi, S. Ratnasamy, and J. Padhye. Hot data centers vs. cool peers. In *Proc. USENIX Workshop on Power aware computing and systems (HotPower)*, December 2008.

[293] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[294] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[295] Dusit Niyato and Ekram Hossain. Dynamics of network selection in heterogeneous wireless networks: an evolutionary game approach. 58(4):2008–2017, 2009.

[296] Mariana Olvera-Cravioto and Octavio Ruiz-Lacedelli. Parallel queues with synchronization. *arXiv preprint arXiv:1501.00186*, 2014.

[297] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose. Modeling TCP Reno performance: a simple model and its empirical validation. 8:133–145, April 2000.

[298] Q. Pang, S. Liew, J. Lee, and V. Leung. Performance evaluation of an adaptive backoff scheme for WLAN. *Wirel. Commun. Mob. Comput.*, 4(8):867–879, December 2004.

[299] H. Park, R. Ratzin, and M. van der Schaar. Peer-to-peer networks – protocols, cooperation and competition. *Streaming Media Architectures, Techniques, and Applications: Recent Advances,* IGI Global, pages 262–294, 2011.

[300] H. Park and M. van der Schaar. Coalition-based resource negotiation for multimedia applications in informationally decentralized networks. 11(4):765–779, June 2009.

[301] B. Peleg and P. Sudhölter. *Introduction to the Theory of Cooperative Games.* Springer-Verlag, 2nd edition, 2007.

[302] Mathew Penrose. *Random geometric graphs.* Number 5. Oxford University Press, 2003.

[303] Mathew D Penrose. On a continuum percolation model. *Advances in applied probability*, 23(03):536–556, 1991.

[304] Mathew D Penrose. Existence and spatial limit theorems for lattice and continuum particle systems. *Prob. Surveys*, 5:1–36, 2008.

[305] Mathew D Penrose. Connectivity of soft random geometric graphs. *The Annals of Applied Probability*, 26(2):986–1028, 2016.

[306] E. Pinheiro, W. Weber, and L. A. Barroso. Failure trends in a large disk drive population. In *Proc. USENIX FAST*, February 2007.

[307] Yuri Pirola, Simone Zaccaria, Riccardo Dondi, Gunnar W Klau, Nadia Pisanti, and Paola Bonizzoni. Hapcol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, 32(11):1610–1617, 2015.

[308] Chris Preston. Spatial birth-and-death processes. Bull. Int. Stat. Inst., Proc. of the 40th Session Warsaw 46, No.2, 371-391 , 1975.

[309] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2):223–252, 1996.

[310] A. Proutiere. Pushing mean field asymptotics to the limits: Stability and performance of random medium access control. In *Proc. Stochastic Networks Conference*, ENS, Paris, June 2008.

[311] A. Proutiere, Y. Yi, T. Lan, and M. Chiang. Resource allocation over network dynamics without timescale separation. In *Proc. IEEE Infocom*, March 2010.

[312] Zrinka Puljiz and Haris Vikalo. Decoding genetic variations: Communications-inspired haplotype assembly. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 13(3):518–530, 2016.

[313] Zrinka Puljiz and Haris Vikalo. Decoding genetic variations: Communications-inspired haplotype assembly. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 13(3):518–530, 2016.

[314] B. Radunović and J.-Y. Le Boudec. Rate performance objectives of multihop wireless networks. 3(4):334–349, October 2004.

[315] B. Radunović and J.-Y. LeBoudec. Power control is not required for wireless networks in the linear regime. In *Proc. IEEE WoWMoM*, Taormina, Italy, June 2005.

[316] B. Raghavan and J. Ma. The energy and emergy of the Internet. In *Proc. ACM Workshop on Hot Topics in Networks (HotNets)*, November 2011.

[317] V. Ramaiyan, A. Kumar, and E. Altman. Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness, multistability and throughput differentiation. In *Proc. ACM Sigmetrics*, June 2005.

[318] V. Ramaiyan, A. Kumar, and E. Altman. Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness, multistability. 16(5):1080–1093, October 2008.

[319] V. Ramaswamy, S. Adlakha, S. Shakkottai, and A. Wierman. Incentives for P2P-assisted content distribution: If you can't beat 'em, join 'em. In *Proc. Allerton Conference on Communication, Control, and Computing,* available at `http://users.cms.caltech.edu/~adamw/papers/p2pincentives.pdf`, October 2012.

[320] S. I. Resnick. *Heavy-Tail Phenomena: Probabilitic and Statistical Modeling.* Springer, 2007.

[321] RNCOS. Global IPTV market forecast to 2014. *Market Research Report*, February 2011.

[322] Philippe Robert. *Stochastic networks and queues.* Springer-Verlag, 2003.

[323] T. Rolski, H. Schmidli, and T. Teugels. *Stochastic Processes for Insurance and Finance.* John Wiley & Sons Inc., 1999.

[324] J. Rosenberg. Interactive connectivity establishment. *RFC (Request For Comments) 5245*, April 2010.

[325] J. Rosenberg, R. Mahy, P. Matthews, and D. Wing. Session traversal utilities for nat. *RFC (Request For Comments) 5389*, April 2008.

[326] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, 3rd edition, 1976.

[327] W. Saad, Z. Han, T. Başar, M. Debbah, and A. Hjørungnes. Hedonic coalition formation for distributed task allocation among wireless agents. accepted for publication in *IEEE Trans. Mobile Comput.*, October 2010.

[328] W. Saad, Z. Han, M. Debbah, A. Hjørungnes, and T. Başar. Coalitional game theory for communication networks. 26(5):77–97, 2009.

[329] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832, 2002.

[330] Shaghayegh Sahebi and William W Cohen. Community-based recommendations: a solution to the cold start problem. In *Workshop on recommender systems and the social web, RSWEB*, 2011.

[331] T. Sakurai and H. Vu. MAC access delay of IEEE 802.11 DCF. 6(5):1702–1710, May 2007.

[332] S.L. Salas and Einar Hille. *Calculus: One and Several Variable.* John Wiley and Sons, New York, 1978.

[333] G. Samorodnitsky. Long-memory and self-similar processes. *Annales de la Faculté des Sciences de Toulouse Mathématiques*, 15(1):107–123, 2006.

[334] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin. Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. In *Proc. ACM MobiCom*, pages 302–314, Philadelphia, PA, September 2004.

[335] Abishek Sankararaman. Spatial stochastic models in wireless and data networks. `http://abishek90.github.io/QualReport.pdf`, May 2017.

[336] Abishek Sankararaman and François Baccelli. Spatial birth–death wireless networks. *IEEE Transactions on Information Theory*, 63(6):3964–3982, 2017.

[337] Abishek Sankararaman and François Baccelli. Community detection on euclidean random graphs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2181–2200. SIAM, 2018.

[338] Abishek Sankararaman, François Baccelli, and Sergey Foss. Interference queueing networks on grids. *Annals of Applied Probability (To Appear)*, 2019. (available at `http://arxiv.org/abs/1710.09797`).

[339] Abishek Sankararaman and Sayan Banerjee. Moment bounds for interference queueing networks. *In Preparation*, 2019.

[340] Abishek Sankararaman, Jeong-woo Cho, and François Baccelli. Performance-oriented association in large cellular networks with technology diversity. In *2016 28th International Teletraffic Congress (ITC 28)*, volume 1, pages 94–102. IEEE, 2016.

[341] Abishek Sankararaman, Jeong woo Cho, and François Baccelli. Performance-oriented association in large cellular networks with technology diversity. *Technical Report*, March 2016. (available at `http://arxiv.org/abs/1603.06928`).

[342] A. Seetharam, M. Somasundaram, D. Towsley, J. Kurose, and P. Shenoy. Shipping to streaming: Is this shift green? In *Proc. ACM Sigcomm Workshop on Green Networking*, August 2010.

[343] S. Sesia, G. Caire, and G. Vivier. Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes. 52:1311–1321, August 2004.

[344] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

[345] Devavrat Shah, NC David, and John N Tsitsiklis. Hardness of low delay network scheduling. *IEEE Transactions on Information Theory*, 57(12):7810–7817, 2011.

[346] Devavrat Shah, Jinwoo Shin, et al. Randomized scheduling algorithm for queueing networks. *The Annals of Applied Probability*, 22(1):128–171, 2012.

[347] Devavrat Shah, Jinwoo Shin, and Prasad Tetali. Efficient distributed medium access. *arXiv preprint arXiv:1104.2380*, 2011.

[348] Virag Shah and Gustavo de Veciana. Performance evaluation and asymptotics for content delivery networks. In *INFOCOM, 2014 Proceedings IEEE*, pages 2607–2615. IEEE, 2014.

[349] A. Shaikh, J. Rexford, and K. Shin. Load-sensitive routing of long-lived IP flows. In *Proc. ACM Sigcomm*, August 1999.

[350] Claude E. Shannon. Two-way communication channels. In *Proc. of the 4th Berkeley Symp. on Mathematical Statistics and Probability, Vol 1*, pages 611–644, Berkeley, Calif., 1961.

[351] L. Shapley. *A Value for n-Person Games.* In H. W. Kuhn and A. W. Tucker, editors, Contribution to the Theory of Games II, vol. 28 of Annals of Mathematics Studies, Princeton University Press, 1953.

[352] G. Sharma, A. Ganesh, and P. Key. Performance analysis of contention based medium access control protocols. 55(4):1665–1681, April 2009.

[353] N. Shephard. From characteristic function to distribution function: a simple framework for the theory. *Econometric theory*, 7(4):519–529, 1991.

[354] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[355] AN Shiryaev. Absolute continuity and singularity of probability measures in functional spaces. In *Proceedings of the International Congress of Mathematicians, Helsinki*, pages 209–225, 1978.

[356] Seva Shneer and Alexander Stolyar. Stability and moment bounds under utility-maximising service allocations, with applications to some infinite networks. *arXiv preprint arXiv:1812.01435*, 2018.

[357] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*, volume 59. Siam, 2009.

[358] G. Simons and N. Johnson. On the convergence of binomial to Poisson distributions. *Ann. Math. Statist.*, 42(5):1735–1736, 1971.

[359] Sushil Singh, Harpreet S Dhillon, and Jeffrey G Andrews. Offloading in heterogeneous networks: Modeling, analysis, and design insights. *Wireless Communications, IEEE Transactions on*, 12(5):2484–2497, 2013.

[360] Rayadurgam Srikant and Lei Ying. *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[361] A. L. Stolyar. On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation. *Oper. Res.*, 53(1):12–25, January 2005.

[362] Alexander L Stolyar. An infinite server system with customer-to-server packing constraints. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1713–1720. IEEE, 2012.

[363] Y. Sun, F. Liu, B. Li, B. Li, and X. Zhang. FS2You: peer-assisted semi-persistent online storage at a large scale. In *Proc. IEEE International Conference on Computer Communications (Infocom)*, April 2009.

[364] V. Suryaprakash, J. Møller, and G. Fettweis. On the modeling and analysis of heterogeneous radio access networks using a Poisson cluster process. 14(2):1035–1047, February 2015.

[365] A.-S. Sznitman. Topics in propagation of chaos. *Lecture Notes in Mathematics,* Ecole d'Etété de Probabilités de Saint-Flour XIX (1989), 1464:165–251, 1991.

[366] F. Tang, L. Deneire, M. Engels, and M. Moonen. A general optimal switching scheme for link adaptation. In *Proc. IEEE VTC Fall*, pages 1598–1602, Atlantic City, NJ, October 2001.

[367] M. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *ACM Comput. Commun. Rev.*,

27(2):5–23, April 1997.

[368] Emre Telatar. Capacity of multi-antenna gaussian channels. *European transactions on telecommunications*, 10(6):585–595, 1999.

[369] O. Tickoo and B. Sikdar. On the impact of IEEE 802.11 MAC on traffic characteristics. 21(2):189–203, February 2003.

[370] O. Tickoo and B. Sikdar. Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks. In *Proc. IEEE Infocom*, March 2004.

[371] Jeffrey Travers and Stanley Milgram. The small world problem. *Phychology Today*, 1(1):61–67, 1967.

[372] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[373] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[374] A. Tutic. The Aumann-Drèze value, the Wiese value, and stability: A note. *International Game Theory Review*, 12(2):189–195, 2010.

[375] V. Valancius, N. Laoutaris, and L. Massoulié. Greening the Internet with nano data centers. In *Proc. ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, December 2008.

[376] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez. Greening the Internet with Nano Data Centers. In *Proc. ACM CoNEXT*, December 2009.

[377] Rahul Vaze. *Random Wireless Networks*. Cambridge University Press, 2015.

[378] A. Veres and M. Boda. The chaotic nature of TCP congestion control. In *Proc. IEEE Infocom*, March 2000.

[379] P. Viswanath, D. N. C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. 48:1277–1294, June 2002.

[380] Nikita Vvedenskaya, Roland Dobrushin, and Fridrikh Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[381] Wei Wang, Xiaobing Wu, Lei Xie, and Sanglu Lu. Femto-matching: Efficient traffic offloading in heterogeneous cellular networks. In *Proc. IEEE Infocom*, 2015.

[382] Harrison C White, Scott A Boorman, and Ronald L Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780, 1976.

[383] W. Whitt. Queues with superposition arrival processes in heavy traffic. *Stochastic Processes and their Applications*, 21:81–91, 1985.

[384] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues.* Springer-Verlag, 2001.

[385] I. Widjaja and A. Elwalid. MATE: MPLS adaptive traffic engineering. IETF Draft, 1999.

[386] W. Willinger, V. Paxson, R. Riedi, and M. Taqqu. Long-range dependence and data network traffic. *Theory and applications of long-range dependence,* Birkhäuser Boston, pages 373–407, 2003.

[387] Ronald W Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[388] Minzhu Xie, Qiong Wu, Jianxin Wang, and Tao Jiang. H-pop and h-popg: Heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics*, 32(24):3735–3744, 2016.

[389] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In *COLT*, pages 903–920, 2014.

[390] X. Yang and G. De Veciana. Service capacity of peer to peer networks. In *Proc. IEEE International Conference on Computer Communications (Infocom)*, March 2004.

[391] Qiaoyang Ye, Beiyu Rong, Yudong Chen, Mazin Al-Shalash, Constantine Caramanis, and Jeffrey G Andrews. User association for load balancing in heterogeneous cellular networks. 12(6), 2013.

[392] Lei Ying, R Srikant, and Xiaohan Kang. The power of slightly more than one sample in randomized load balancing. *Mathematics of Operations Research*, 42(3):692–722, 2017.

[393] H. Zhai, Y. Kwon, and Y. Fang. Performance analysis of IEEE 802.11 MAC protocols. *Wirel. Commun. Mob. Comput.*, (4):917–931, 2004.

[394] B. Zhang, S. Borst, and M. Reiman. Optimal server scheduling in hybrid P2P networks. *Performance Evaluation*, 67:1259–1272, 2010.

[395] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok. A novel ultrathin elevated channel low-temperature poly-Si TFT. 20:569–571, November 1999.

[396] H. Zheng and H. Viswanathan. Optimizing the ARQ performance in downlink packet data systems with scheduling. 4:495–506, March 2005.

[397] Yu Zheng. Tutorial on location-based social networks. WWW 2012, May 2012.

[398] J. Zhu and B. Hajek. Stabilty of peer to peer systems. In *Proc. ACM Symposium on Principles of Distributed Computing (PODC)*, June 2011.

[399] M. Zorzi and R. R. Rao. On the use of renewal theory in the analysis of ARQ protocols. 44:1077–1081, September 1996.

[400] AP Zwart and Onno J. Boxma. Sojourn time asymptotics in the m/g/1 processor sharing queue. *Queueing systems*, 35(1-4):141–166, 2000.

Vita

Abishek Sankararaman was born in New Delhi, India on 24 March 1990. He received the Bachelor of Technology and Masters of Technology degrees, both in Electrical Engineering from the Indian Institute of Technology Madras, Chennai in India in 2013. Since August 2013, he has been at The University of Texas at Austin, where he is a graduate student in the Department of Electrical and Computer Engineering. In the summer of 2015, he spent three months as a data-science intern in Huawei Research Labs in Santa Clara, CA.

Permanent address: abishek.90@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.