Fundamental Limits and Optimal Operation in Large Wireless Networks

THÈSE N^{O} 4483 (2009)

PRÉSENTÉE LE 30 OCTOBRE 2009

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE THÉORIE DE L'INFORMATION
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Ayfer ÖZGÜR AYDIN

acceptée sur proposition du jury:

Prof. R. Urbanke, président du jury
Dr O. Lévêque, Prof. E. Telatar, directeurs de thèse
Prof. H. Bölcskei, rapporteur
Prof. S. Shamai (Shitz), rapporteur
Prof. D. Tse, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2009

Anneme ve Babama

Abstract

Wireless adhoc networks consist of users that want to communicate with each other over a shared wireless medium. The users have transmitting and receiving capabilities but there is no additional infrastructure for assisting communication. This is in contrast to existing wireless systems, cellular networks for example, where communication between wireless users heavily relies on an additional infrastructure of base stations connected with a high-capacity wired backbone. The fact that they are infrastructureless makes wireless adhoc networks inexpensive, easy to build and robust but at the same time technically more challenging. The fundamental challenge is how to deal with interference: many simultaneous transmissions have to be accommodated on the same wireless channel when each of these transmissions constitutes interference for the others, degrading the quality of the communication.

The traditional approach to wireless adhoc networks is to organize users so that they relay information for each other in a multi-hop fashion. Such multi-hopping strategies face scalability problems at large system size. As shown by Gupta and Kumar in their seminal work in 2000, the maximal communication rate per user under such strategies scales inversely proportional to the square root of the number of users in the network, hence decreases to zero with increasing system size. This limitation is due to interference that precludes having many simultaneous point-to-point transmissions inside the network.

In this thesis, we propose a multiscale hierarchical cooperation architecture for distributed MIMO communication in wireless adhoc networks. This novel architecture removes the interference limitation at least as far as scaling is concerned: we show that the per-user communication rate under this strategy does not degrade significantly even if there are more and more users entering into the network. This is in sharp contrast to the performance achieved by the classical multi-hopping schemes.

However, the overall picture is much richer than what can be depicted by a single scheme or a single scaling law formula. Nowadays, wireless adhoc networks are considered for a wide range of practical applications and this translates to having a number of system parameters (e.g., area, power, bandwidth) with large operational range. Different applications lie in different parameter ranges and can therefore exhibit different characteristics. A thorough under-

standing of wireless adhoc networks can only be obtained by exploring the whole parameter space. Existing scaling law formulations are insufficient for this purpose as they concentrate on very small subsets of the system parameters. We propose a new scaling law formulation for wireless adhoc networks that serves as a mathematical tool to characterize their fundamental operating regimes.

For the standard wireless channel model where signals are subject to power path-loss attenuation and random phase changes, we identify four qualitatively different operating regimes in wireless adhoc networks with large number of users. In each of these regimes, we characterize the dependance of the capacity on major system parameters. In particular, we clarify the impact of the power and bandwidth limitations on performance. This is done by deriving upper bounds on the information theoretic capacity of wireless adhoc networks in Chapter 3, and constructing communication schemes that achieve these upper bounds in Chapter 4. Our analysis identifies three engineering quantities that together determine the operating regime of a given wireless network: the short-distance signal-to-noise power ratio (SNR_s), the long-distance signal-to-noise power ratio (SNR_l) and the power path-loss exponent of the environment. The right communication strategy for a given application is dictated by its operating regime. We show that conventional multi-hopping schemes are optimal when the power path-loss exponent of the environment is larger than 3 and $\mathrm{SNR}_s \ll 0$ dB. Such networks are extremely power-limited. On the other hand, the novel architecture proposed in this thesis, based on hierarchical cooperation and distributed MIMO, is the fundamentally right strategy for wireless networks with $\mathrm{SNR}_l \gg 0$ dB. Such networks experience no power limitation. In the intermediate cases, captured by the remaining two operating regimes, neither multi-hopping nor hierarchical-MIMO achieves optimal performance. We construct new schemes for these regimes that achieve capacity.

The proposed characterization of wireless adhoc networks in terms of their fundamental operating regimes, is analogous to the familiar understanding of the two operating regimes of the point-to-point additive white Gaussian noise (AWGN) channel. From an engineering point of view, one of the most important contributions of Shannon's celebrated capacity formula is to identify two qualitatively different operating regimes on this channel. Determined by its signal-to-noise power ratio (SNR), an AWGN channel can be either in a bandwidth-limited ($\mathrm{SNR} \gg 0$ dB) or a power-limited ($\mathrm{SNR} \ll 0$ dB) regime. Communication system design for this channel has been primarily driven by the operating regime one is in.

Keywords: Wireless Adhoc Networks, Scaling Laws, Linear Scaling, Capacity of Wireless Networks, Hierarchical Cooperation, Multi-hopping, Distributed MIMO, Operating Regimes, Throughput-Delay Tradeoff, Random Matrix Theory

Résumé

Les réseaux ad hoc sans fil sont composés d'utilisateurs qui désirent communiquer entre eux sur une même bande de fréquence. Chaque utilisateur est capable d'émettre et de recevoir, mais il n'y a pas d'infrastructure fixe pour relayer les communications. Ces réseaux diffèrent donc des systèmes sans fil existants, comme les réseaux cellulaires par exemple, dans lesquels les communications entre utilisateurs sont relayées par des stations de base reliées entre elles par un réseau câblé. L'absence d'infrastructure dans les réseaux ad hoc présente des avantages en termes de coût, de fabrication et de robustesse, mais représente en même temps un défi du point de vue technologique. Le plus important de ces défis concerne la bonne gestion des interférences: plusieurs communications doivent être relayées simultanément sur une même bande de fréquence, alors que chaque communication constitue de l'interférence pour les autres, contribuant à dégrader la qualité des transmissions.

L'approche traditionnelle des réseaux ad hoc sans fil consiste à organiser les utilisateurs de sorte à ce qu'ils relayent de proche en proche l'information destinée aux uns et aux autres. De telles stratégies ne fonctionnent malheureusement pas à grande échelle. Ce fait a été mis en évidence par le travail de Gupta et Kumar en 2000: plus précisément, il a été établi qu'avec de telles stratégies, le taux maximal de communication par utilisateur décroît comme l'inverse de la racine carrée du nombre d'utilisateurs dans le réseau, et donc tend vers zéro lorsque le système grandit. Cette limitation provient de l'interférence qui empêche d'effectuer plusieurs transmissions simultanées d'utilisateur à utilisateur dans le réseau.

Dans la présente thèse, nous proposons un nouvelle stratégie multi-échelle pour établir des communications multi-utilisateurs distribuées dans un réseau ad hoc sans fil. Cette nouvelle architecture permet de supprimer (en termes de loi d'échelle) la limitation imposée par l'interférence: nous montrons en effet qu'en utilisant cette stratégie, le taux de communication par utilisateur ne se dégrade pas significativement, même lorsque le nombre d'utilisateurs augmente dans le réseau. Ce résultat contraste donc fortement avec la performance des schémas traditionnels évoqués plus haut.

Cependant, l'étude de la capacité des réseaux ad hoc sans fil ne se résume pas à un seul schéma de communication, ni à une seule loi d'échelle. De nom-

breuses applications où de tels réseaux démontreraient leur utilité sont envisagées actuellement, ce qui se traduit par des grandes variations des valeurs des paramètres caractérisant le système (puissance, largeur de bande, aire du réseau, etc.). En fonction de l'application considérée, le système peut exhiber des caractéristiques très différentes. Une description complète du comportement de ces réseaux passe donc par l'exploration de tout l'espace des paramètres du système. Les lois d'échelle existant dans la littérature ne permettent pas d'obtenir une telle description, car elles ne se concentrent que sur une petite fraction de l'espace des paramètres. Nous proposons une nouvelle formulation qui permet de caractériser les régimes opératoires fondamentaux des réseaux ad hoc sans fil.

Pour le modèle standard de transmission sans fil, dans lequel on considère que l'amplitude des signaux décroît avec la distance en loi de puissance et que les changements de phase sont aléatoires, nous identifions quatre régimes opératoires qualitativement différents. Dans chacun de ces régimes, nous caractérisons la dépendance de la capacité dans les principaux paramètres du système. En particulier, nous clarifions l'impact des limitations de puissance et de largeur de bande sur la performance. Pour ce faire, nous dérivons des bornes supérieures sur la capacité des réseaux ad hoc sans fil dans le chapitre 3; puis nous décrivons au chapitre 4 des schémas de communication qui permettent d'atteindre ces bornes supérieures (en termes de loi déchelle). Notre analyse permet d'identifier trois quantités qui déterminent les différents régimes opératoires d'un réseau: le rapport signal sur bruit à courte distance (SNR_s), le rapport signal sur bruit à longue distance (SNR_l) et l'exposant d'atténuation des signaux avec la distance (α). Le bon schéma de communication à adopter est dicté par le régime sans lequel on se trouve. Nous montrons que les schémas de communication traditionnels (relayage des communications de proche en proche) sont optimaux lorsque $\alpha \geq 3$ et $\mathrm{SNR}_s \ll 0$ dB. De tels réseaux sont fortement limités en puissance. A l'autre extrême, le nouveau schéma de communication proposé dans cette thèse, basé sur une coopération hiérarchique et des communications multi-utilisateurs distribuées, est la bonne stratégie à adopter dans les réseaux ad hoc lorsque $\mathrm{SNR}_l \gg 0$ dB. De tels réseaux ne sont pas limités en puissance. Pour les cas intermédiaires, qui correspondent aux deux régimes restants, aucune des deux stratégies mentionnées ci-dessus n'est optimale. Nous construisons de nouveaux schémas de communication qui atteignent la capacité dans ces deux cas.

La caractérisation des réseaux ad hoc sans fil proposée ici en termes de ses quatre régimes opératoires fondamentaux est analogue à la description familière des deux régimes opératoires du canal avec bruit blanc gaussien additif (AWGN). Du point de vue de l'ingénieur, l'une des contributions les plus importantes du célèbre théorème de Shannon sur la capacité est d'identifier deux régimes qualitativement différents pour ce canal. En fonction de son rapport signal sur bruit (SNR), un canal AWGN est limité soit en fréquence (si $\mathrm{SNR} \gg 0$), soit en puissance (si $\mathrm{SNR} \ll 0$). Le design de schémas de communication pour ce canal est déterminé en prioriété par le régime opératoire

dans lequel on se trouve.

Mots-clés: réseaux sans fil “ad hoc”, lois déchelle, capacité de reśeaux sans fil, comportement linéaire, coopération hiérarchique, communications avec relais multiples, systèmes multi-antennes distribués, régimes opératoires, compromis délai-débit.

Acknowledgments

I feel honored to hold the distinguished title of being the first and (till now) only PhD student of Dr. Olivier Lévêque. This title allowed me to enjoy a number of privileges, which I think very few PhD students can enjoy. "Do you have a minute, I have an idea?" or more often "Do you have a minute, I am miserably confused?": It was never "a minute" but the answer was always "yes". Olivier always had the time and patience to listen to my vague ideas and to prove (and re-prove) matrix inequalities for me. I am indeed grateful to him not only for listening to my ideas, but for taking them seriously each time, thinking together with me and sharing the excitement and hope until these ideas proved to be useless as they usually did. He has been an advisor with a genuine concern for my welfare, an inspiring collaborator, a personal math guru, and a great friend. For all of the above and many other reasons, I want to express my deepest gratitude to him. When my future students (if any) knock unexpectedly on my door, I will try to remember how much patience and generosity I received from Olivier.

Prof. Emre Telatar has been the first person to look for, whenever I needed an advice on almost anything. This is not only because Emre is my coadvisor, but rather because of his well-known wisdom, kindness and generosity of heart. I enjoyed a large number of discussions with him, sometimes on research, sometimes on personal stuff, which enriched me in many ways. I am deeply indebted to him for his critical help in a number of issues related to my stay in Lausanne. Without his help, my stay here would not have been comfortable, if at all possible.

I was very fortunate to have the opportunity to collaborate with Prof. David Tse during my PhD study. This thesis would not have been possible had Olivier not forwarded an e-mail of mine to David three years ago, which started our collaboration. I had an amazing research experience with David and Olivier, almost all of which was long-distance. To give an example, I have just counted forty-seven e-mails exchanged among the three of us on the day of August 1, 2006, which were most likely accompanied by a number of skype meetings. The contents of these e-mails form the pages of the following thesis. However, I am most grateful to David not for what goes into these pages, but for what I think really qualifies one for the title of *Philosophiæ Doctor*: I hope,

I have learned how to do research from David and, just as importantly, how to enjoy it. I am extremely thankful to him for introducing me to his unique style of research and for sharing his insight, experience, and perspective on the field. My collaboration experience with David has mostly shaped the way I view research and the fields of information theory and wireless communication today.

I would like to thank my thesis committee, Prof. Shlomo Shamai and Prof. Helmut Bölcskei for reading my thesis and their valuable comments on my research. I also would like to thank Prof. Rüdiger Urbanke for agreeing to be the president of my committee and for his interest in my career. I am also grateful to Prof. Suhas Diggavi whose graduate class on wireless networks I followed three times. Every time it was an inspiring and insightful experience. I had the chance to benefit from various discussions with Prof. Bixio Rimoldi, Prof. Christina Fragouli and Prof. Patrick Thiran. Finally, I would like to thank Prof. Ramesh Johari, with whom I had the chance to collaborate towards the end of my PhD.

Life in EPFL would have been much more difficult without the help of two people, Yvonne Huskie and Damir Laurenzi. The workload of these two people will probably decrease by a great deal once I leave the lab. Yvonne had to prepare four contracts for me in the last four months. On the other hand, I always had a question (sometimes an embarrassingly silly one) for Damir about computers, each time he would pass in front of my office. I am grateful to these people as well as Muriel and Françoise for the welcoming smile whenever I showed up at their door. In this line, I should also thank "the husband", as everyone in our group calls him, for his service as my system administrator at home.

There are many friends who enriched my life at EPFL. One of them is my dear officemate, Marius Kleiner. His "random talks", usually shocking jokes, computer and English skills, creative mind and liberal thoughts will be sorely missed. Without listing names and taking the risk to forget someone, I would also like to thank all my colleagues in IPG, all of them great friends, for various discussions on research and the good times we shared.

Last but not least, I would like to thank my family. I would like to thank my parents-in-law for their constant support and encouragement during our stay in Lausanne. I would like to thank my brother Ayhan for trying to find words to encourage me during difficult times. It feels strange to formally thank my parents as their love and support has always felt so natural, taken for granted. Moreover, there are no words that can fully express my gratitude to them. As a futile attempt, I dedicate this thesis to my parents.

Lastly, I would like to express my most heartfelt gratitude to my husband, Bilge. After sharing everything during the time of this PhD study — our days, our responsibilities, our successes and failures, our joys and sorrows — it is not very meaningful to talk about a thesis or an achievement that belongs to only one of us. Even though Bilge's first name does not appear on the cover page, it is more reasonable to consider this thesis as a joint work of ours.

This thesis was supported in part by Swiss NSF grants Nr 200021-108089 and 200020-118076.

Contents

1 Introduction

The last decade has witnessed the rise of wireless technology in everyday life. The most prominent examples of this technology are cellular networks and wireless LANs (local area networks), which have made cellular phones and laptops our daily companions. The rise in practice has been preceded by a surge of research activity in wireless communication theory, which has led to an understanding of the fundamental trade-offs involved in the design of such systems and the possible techniques to achieve high performance. This understanding has enabled the design of wireless systems offering comparable quality of service to their wireline counterparts. With the inherent advantage of mobility and ease of deployment, these wireless systems have become more popular today than their wireline alternatives.

Nevertheless, the role of wireless technology in current communication services remains still very limited. In cellular networks and wireless LANs, the wireless system provides only the last stage of communication, from the so called base stations (in cellular networks) or access points (in wireless LAN) to the end users. The communication between the base stations or access points is carried by wired high-capacity links. The need to install an extensive infrastructure of base stations, access points and a high capacity backbone, makes these systems expensive, difficult to build and not robust enough for certain applications.

Why not get rid of the heavy burden of installing an expensive infrastructure and just let users communicate among themselves? That is the basic motivation for the so called *wireless adhoc networks* studied in this thesis. Wireless adhoc networks differ from the conventional infrastructure-based networks above by the fact that they rely completely on wireless communication. They are simply formed by a group of users, usually called nodes, that have transmitting and receiving capabilities. The nodes can be the mobile phones

of the cellular topology, laptops like in WLANs, or sensors that measure some physical data. Whatever the application is, the common characteristic is the following: A group of nodes want to communicate with each other over the shared wireless medium but there is no additional infrastructure for assisting communication or for coordinating traffic. The users need to self-organize and relay information for each other while all communication must happen exclusively over the wireless channel.

Obviously, designing such stand-alone wireless networks is technically much more challenging than designing networks that can rely heavily on an additional infrastructure for coordination and communication. First of all, the *adhoc* nature of these networks gives rise to a number of unique challenges in coordinating the nodes in a decentralized fashion. However, the *all-wireless* nature of the network poses a much more fundamental challenge. Even if the nodes were able to coordinate and follow such a strategy, is there any efficient strategy at all for wireless adhoc networks? The challenge is that many simultaneous transmissions have to be accommodated on the same wireless channel when each of these transmissions constitutes *interference* for the others, degrading the quality of the communication.

Indeed, the phenomenon of interference lies at the heart of every multi-user wireless communication problem. However, in infrastructure-based networks, this problem is circumvented by constraining wireless communication to be only of *local* nature. In such networks, the gateways to the wired backbone (base stations or access points) are spread densely over the network area so that wireless users have to communicate only to a nearby gateway. Since wireless signals get attenuated over distance, many simultaneous *local* communications can be established over the same wireless channel. In other words, gateways that are sufficiently separated in space can serve their users simultaneously on the same wireless channel without creating too much interference for each other. In case there are multiple users that want to access the same gateway, resources are shared among users. More precisely, signaling dimensions such as time, frequency and code space are divided between users so that communications with different users are orthogonal to each other, i.e., they do not interfere. If the number of users served by a gateway is small, which is ensured by the dense installation of gateways over the network area, this simple strategy of sharing resources yields acceptable performance.

The problem of interference poses a greater challenge in wireless adhoc networks. Typically, the traffic requirements of the network are such that there are many *long-distance* communications that need to be established over the same wireless channel. Dividing resources among users is not a good strategy anymore. Resources are scarce and simply dividing them among the many users in the network leads poor performance. Innovative approaches to interference management are required in order to design efficient and high performance wireless adhoc networks. On the other hand, *long-distance* wireless communication is also challenging from the *power* point of view. Even if there were no interference, direct communication between two distant users might not be

possible. Wireless signals get attenuated over distance and may not be able to reach a destination far away with sufficient power.

Given these technical challenges, which seem unbearable at first sight, one may be surprised to see how much interest wireless adhoc networks have attracted from the networking, communication and information theory communities over the last 15 years. They have been subject to intensive research, leading to an increasing number of conferences, journals and books specializing particularly on this topic. The motivation for the persistent research, despite the technical challenge, comes from a number of critical advantages offered by these networks. The fact they do not require any infrastructure makes them inexpensive, easy to build and robust. They can be incorporated inside the existing cellular or WLAN topologies for assisting communication in a cost-efficient way. They are also envisaged to enable a new collection of exciting wireless applications in near future. Such applications include emergency and military applications, sensor networks, vehicular communication, smart homes, etc.

Before these exciting ideas become reality, however, a number of critical open problems need to be resolved. One of the most important open problems in this field is to develop a fundamental understanding of these networks from a capacity point of view. Given a particular wireless adhoc network topology, what is the best communication rate we can get for every user in this network? How does this rate depend on system parameters like the total number of users inside the network, the power budget per user, the total bandwidth allocated for communication, the area of the network etc.? How should we design our transmitters and receivers and how should we organize communications to approach this ultimate limit? How do different strategies compare to each other and what are the right strategies for different wireless adhoc network applications? An understanding of the capacity of wireless adhoc networks will yield critical intuition about the answer of such fundamental engineering questions. It will provide operating principles and architectural guidelines for the design and deployment of wireless adhoc systems. The aim of the current dissertation is to contribute in this direction.

1.1 Current vs. Future Wireless Technology

From the point of view of this dissertation, the literature on wireless adhoc networks can be divided into two main groups. The first group reflects the "networking" approach to wireless adhoc networks. This line of research studies wireless adhoc networks in the light of the theories and methods developed for traditional wired networks. Roughly speaking, the aim here is to mimic the operation of a wired network inside a wireless network. As wired networks form graphs, packets are relayed from one node on the graph to another by multi-hopping on a path connecting these two nodes. At each hop, the relay nodes fully decode the transmitted packets, re-encode and forward them to

the next node on the path. The "networking" approach suggests to apply the same multi-hopping strategy in wireless adhoc networks. However, there are a number of challenges due to fundamental differences between wireless and wired networks. The first obvious difficulty is that while a wired network naturally induces a graph, there is no graph corresponding to a wireless network. Hence, one needs to start by artificially associating a graph with the wireless network. The common approach is to draw a connectivity graph. A link between two nodes on this graph indicates that these nodes are within the transmission range of each other. However, this graph is not unique as in the wired case. For the same physical placement of nodes, the connectivity graph of a wireless adhoc network can look very different under different choices of operational power and rate. Moreover, the links in this graph are not independent. The links that are sufficiently separated in space can be activated at the same time, but there is interference between neighboring links. The main focus of this line of research is to overcome such additional challenges introduced by the wireless and also decentralized nature of the setup and extend the well-established techniques in medium access control, routing, etc., for wired multi-hop networks, to wireless networks.

The "networking" approach is also motivated by the state-of-the-art physical-layer wireless technology. Although a number of multi-user techniques are known for wireless communication, current wireless systems are mostly restricted to point-to-point communication. Usually, the aim is to replace the wire between a source point and a destination point by providing a wireless link of certain capacity between them. Motivated by this fact, the "networking" approach views the wireless network as a collection of such point-to-point links, which yields a connectivity graph as discussed earlier. However, this is a very restrictive treatment of wireless adhoc networks. A wireless network is inherently much more complex than a collection of point-to-point links and indeed this complex structure opens many other possibilities for enhancing information transfer. The fact that point-to-point communication with single-user encoding and decoding of messages is prominent in current wireless systems does not necessarily imply that future wireless systems should confine to the same physical-layer technology. Principles rooted in the current engineering practice may not be applicable in general, and can lead to very suboptimal designs. This is the motivation for the second group of research, that can be called the "wireless" approach, since the emphasis now is on the wireless nature of the problem. This approach views wireless adhoc networks as a brand new multi-user wireless communication problem and seeks specific solutions for them. It is rooted in network information theory and is often called the information theoretic approach, as it studies wireless adhoc networks without making any a priori assumption on how they are to operate.

Of special interest to the current dissertation is the the seminal work of P. Gupta and P. R. Kumar in 2000 [27]. This work has introduced a simple yet insightful model for wireless adhoc networks with a large number of users.

The model explicitly takes into consideration important features of wireless networks, such as the spatial distribution of nodes and the traffic requirement between them as well as the attenuation of wireless signals with distance and the broadcasting and superposition nature of wireless media. Using this model, Gupta and Kumar have characterized the potentials and the limitations of the "networking" approach in large wireless networks. The result has also raised interest in the wireless research community as it naturally leads to the question of whether there are more efficient strategies to operate wireless networks. Next, we discuss the results of this work in more detail, as well as the random network model it introduces, as this model forms the starting point for the current thesis. A more detailed overview of the literature on wireless adhoc networks is given in the beginning of each chapter. A general overview of the subject can be found in the survey paper [50].

A Random Network Model

As wireless network applications become large, complex and diverse, there is a need to derive fundamental operating principles that can serve as a rule of thumb in their design and deployment. This requires a fundamental understanding that can not be attained by studying particular instances of such networks. An abstraction is required that captures the essential aspects of the problem while stripping out the less important details. Such an abstract model has been proposed by Gupta and Kumar in [27] that has turned out be tractable and yet useful.

Instead of worrying about the exact placement of nodes inside the network, Gupta and Kumar propose to consider a random model where n nodes are randomly distributed over a two dimensional area of unit size. The traffic is also random: Every node is both a source and a destination and the sources and destinations are randomly paired up one-to-one without any consideration on node locations. Each source has the same traffic rate R to send to its destination node. On the physical side, signals transmitted from one user to another at distance r apart are assumed to experience a power loss of $r^{-\alpha}$ and a random rotation in the phase.[1] The parameter $\alpha \geq 2$ is called the power path loss exponent of the environment, $\alpha = 2$ corresponding to free space propagation. Every user has a fixed power budget of P Watts, and the wireless system is allocated a fixed bandwidth of W Hz. The question of interest is the maximally achievable total throughput $T = nR$ in such a network. In order to make the problem tractable, Gupta and Kumar restrict attention to the scaling of this maximally achievable total throughput $T(n)$ with increasing system size n. Such a scaling law formulation puts the emphasis on large system size and can be used to understand the behavior of large wireless adhoc networks.

[1] The conclusions of [27] hold regardless of whether the channel model includes a random phase rotation or not.

Gupta and Kumar use the above random network model to study the performance of classical multi-hop architectures in wireless adhoc networks. They show that such strategies based on single-user encoding-decoding and forwarding of packets cannot achieve a total throughput scaling better than $O(\sqrt{n})$ and this maximal scaling can only be achieved if such strategies confine transmissions to take place between nearest neighbors. Single user encoding-decoding implies that the signals received from nodes other than one particular transmitter are interference and should be regarded as noise, degrading the quality of the communication. Given this assumption, long-range communication between source and destination pairs is not a good idea, as the interference generated would preclude most of the other nodes from communicating. Instead, the optimal strategy is to communicate between nearest neighbors and maximize the number of simultaneous transmissions (spatial reuse). However, this means that each packet has to be retransmitted many times before getting to its final destination, leading to a sub-linear scaling of the system throughput.

A total throughput of $O(\sqrt{n})$ implies that the rate $R(n)$ per source-destination pair has to decrease to zero as $O(1/\sqrt{n})$ when the system size n is large. Therefore, this result casts doubt on the feasibility of multi-hop wireless networks on a large scale. The $O(\sqrt{n})$ limitation is the consequence of the interference in the wireless media that precludes source nodes to simply transmit their messages simultaneously to their destination nodes. The conclusions of the work [27] lead to the question of whether this interference barrier can be surpassed with more complex communication strategies. The present dissertation answers this question in the affirmative.

1.1.1 A Hierarchical Cooperation Scheme

One of the main results of this thesis is that one can in fact achieve arbitrarily close to *linear* total throughput scaling: for any $\epsilon > 0$, we present a scheme that achieves an aggregate rate of $\Theta(n^{1-\epsilon})$. This is a surprising result: a linear scaling means that there is essentially *no* interference limitation; the rate for *each* source-destination pair does not degrade significantly, even as one puts more and more nodes in the network. Using tools from information theory, it is easy to show that one cannot get a better capacity scaling than $O(n \log n)$, so the suggested scheme is very close to optimal.

To achieve linear scaling, one must be able to perform *many* simultaneous long-range communications. A physical-layer technique achieving this is MIMO (multi-input multi-output), i.e., the use of multiple transmit and receive antennas to multiplex several streams of data and transmit them simultaneously. MIMO was originally developed in the point-to-point setting, where the transmit antennas are co-located at a single transmit node, each transmitting one data stream, and the receive antennas are co-located at a single receive node, jointly processing the vector of received observations at the antennas. A natural approach to apply this concept to the network setting is to have both source nodes and destination nodes cooperate in *clusters* to form distributed

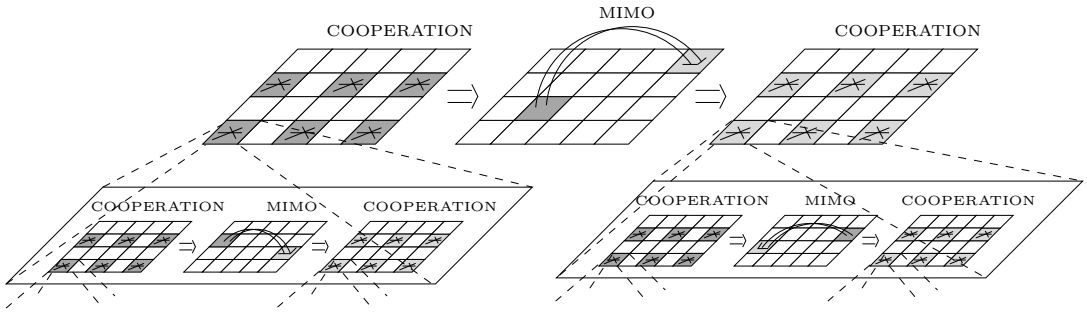

Figure 1.1: The figure illustrates the salient features of the three phase hierarchical scheme in Section 4.4.

transmit and receive antenna arrays, respectively. This way, mutually interfering signals can be turned into useful ones that can be jointly decoded at the receive cluster and spatial multiplexing gain can be realized. In fact, if *all* the nodes in the network could cooperate for free, then a classical MIMO result [17, 49] says that a sum rate scaling proportional to n could be achieved. However, this may be over-optimistic : communication between nodes is required to set up the cooperation and this may drastically reduce the useful throughput. The contribution of this dissertation is to introduce a new *multi-scale, hierarchical* cooperation architecture that does not introduce significant overhead to communication. Such cooperation first takes place between nodes within small local clusters that can operate simultaneously, since the decay of signals with distance allows simultaneous local communications. The cooperation facilitates MIMO communication over a larger spatial scale. This can then be used as a communication infrastructure for cooperation within larger clusters at the next level of the hierarchy. Continuing in this fashion, cooperation can be achieved at an almost global scale. At the highest level of the hierarchy, long-range MIMO communications can be performed between clusters almost as large as the whole network. By increasing the number of levels in such a hierarchical architecture, one can get arbitrarily close to linear aggregate throughput scaling. Figure 1.1 illustrates the hierarchical architecture with a focus on the top two levels.

1.2 A New Scaling Law Formulation for Wireless Ad-hoc Networks

The classical capacity formula

$$C_{AWGN}(W, P_r/N_0) = W \log_2 \left(1 + \frac{P_r}{N_0 W}\right) \quad \text{bits/s} \tag{1.1}$$

of a point-to-point additive white Gaussian noise (AWGN) channel with bandwidth W Hz, received power P_r Watts, and white noise with power spectral

density $N_0/2$ Watts/Hz plays a central role in communication system design. The formula not only quantifies exactly the performance limit of communication in terms of system parameters, but perhaps more importantly also identifies two qualitatively different operating regimes depending on the signal-to-noise power ratio

$$\text{SNR} := \frac{P_r}{N_0 W}.$$

In the *power-limited* (or low SNR) regime, where SNR $\ll 0$ dB, the capacity is approximately linear in the power and the performance depends critically on the power available, but not so much on the bandwidth. In the *bandwidth-limited* (or high SNR) regime, where SNR $\gg 0$ dB, the capacity is approximately linear in the bandwidth and the performance depends critically on the bandwidth, but not so much on the power. This understanding of the two operating regimes of the AWGN channel can be summarized by the following approximation formula for the capacity

$$C_{AWGN}(W, P_r/N_0) \propto \begin{cases} W & \text{SNR} \gg 0 \text{ dB} \\ P_r/N_0 & \text{SNR} \ll 0 \text{ dB}. \end{cases} \tag{1.2}$$

The design of good communication schemes is primarily driven by the operating regime one is in.

Now, imagine the capacity formula (1.1) were not at our disposal and we were interested in finding (1.2) that approximates the dependence of the capacity to the two resources in the channel and identifies two different operating regimes depending on SNR. The approximation (1.2) can be obtained by studying the interplay between the two resources, the bandwidth and the power. Suppose P_r/N_0 and W are coupled to each other via a parametric formula, $P_r/N_0 = W_0 m^{\gamma_1}$ and $W = W_1 m^{\gamma_2}$ with γ_1, γ_2 fixed real numbers and m the dummy parameter. W_0 and W_1 are positive constants of appropriate units. Assume further that for any γ_1, γ_2, we are able to characterize the scaling exponent of the spectral efficiency $\rho_{AWGN} = C_{AWGN}/W$, in bits/s/Hz, with m,

$$e_{AWGN}(\gamma_1, \gamma_2) := \lim_{m \to \infty} \frac{\log \rho_{AWGN}(\gamma_1, \gamma_2)}{\log m}.$$

For the AWGN, we would find

$$e_{AWGN}(\gamma_1, \gamma_2) = \begin{cases} 0 & \gamma_1 - \gamma_2 \geq 0 \\ \gamma_1 - \gamma_2 & \gamma_1 - \gamma_2 < 0. \end{cases}.$$

The above expression can be written in a simpler form,

$$e_{AWGN}(\gamma) = \begin{cases} 0 & \gamma \geq 0 \\ \gamma & \gamma < 0. \end{cases}.$$

if we define $\gamma = \gamma_1 - \gamma_2$ and $\text{SNR} = P_r/N_0 W = m^{\gamma}$. (From now on, we ignore the constants W_0, W_1 that are required for matching the units in the parametric

formula above, but do not change the scaling law.) This characterization of the scaling exponent can be used to deduce the approximation (1.2) for the capacity. Note that the scaling law is not of interest in its own right here. It is used as a tool to discover the operating regimes of the AWGN channel. Even though (1.2) is obtained from a scaling law analysis, (1.2) itself provides an approximation for the capacity of any given AWGN channel. One of the main contributions of this dissertation is to provide an analogous approximation for the capacity of wireless adhoc networks with a large number of users.

The literature on scaling laws for wireless networks has exclusively concentrated on two particular models of how the geometry of the network changes with increasing number of nodes. The first one is the *dense scaling* introduced in Section 1.1. It assumes that the n users are distributed on a unit area, $A = 1$, that remains constant as the number of users in the network increases. The second scaling that has been studied in the literature is the *extended scaling*. In this case, the area of the network extends linearly with increasing number of users, $A = n$, while the rest of the assumptions are same as in the dense scaling. A major effort has been made to characterize the capacity scaling of dense and extended wireless networks in the literature. The extended scaling has been characterized for $\alpha \geq 4$, while the dense scaling has been open. A detailed overview of the literature can be found in the beginning of Chapters 3 and 4.

In this dissertation, we give a complete characterization of both the dense and the extended scalings. In the previous section, we have already discussed that the total capacity scales like $\Theta(n)$ with the number of users n in the dense scaling. This linear scaling is achieved with hierarchical cooperation and MIMO communication. This is not the case in extended scaling. In Chapters 3 and 4, we will show that the capacity scaling of extended networks exhibits a dichotomy depending on the power path loss exponent of the environment. When $2 \leq \alpha < 3$, the capacity scales like $\Theta(n^{2-\alpha/2})$ and this scaling is achieved by the hierarchical cooperation scheme introduced in the previous section. When $\alpha \geq 3$, the capacity scales like $\Theta(\sqrt{n})$ and the optimal strategy is to confine to nearest neighbor point-to-point transmissions and relay packets through multi-hopping inside the network.

Given these two different scaling law results, what can be inferred about a particular wireless network of interest with given number of users (large but finite), area, power and bandwidth budgets, path loss exponent, etc.? Assume for example that $\alpha = 4$. What is the right strategy to operate this wireless network, hierarchical cooperation or multi-hopping? The difficulty is that we do not know which scaling law result, dense or extended, we should take as a basis to design our network. Recall that the motivation for studying this rather abstract formulation of scaling laws for wireless networks was to develop an intuition about the answers to such fundamental engineering questions. The only intuition suggested by these two qualitatively different results obtained for

two different couplings of the system parameters is that the right strategies can be different for different wireless networks. In other words, these two different scaling law results hint the existence of two different operating regimes in wireless networks.

However, the picture is far from complete for wireless networks: First of all, it is not obvious whether there are only two operating regimes in wireless networks. There can be some other operating regimes exhibiting a completely different behavior with neither of the above schemes, hierarchical cooperation or multi-hopping, achieving optimal capacity scaling. Moreover, it is not clear what engineering quantities determine the operating regime a wireless network is in. The characterization of two particular scalings that couple the system resources in two particular ways, does not suffice to develop a comprehensive understanding of wireless networks.

This fact should not be surprising, considering the diversity of wireless adhoc network applications. Wireless networks have a number of parameters with large range that can lead to different behaviors in different applications. With the scaling law formulation, we restrict attention on networks with large number of users, but the area, power and bandwidth of such large networks can still be diverse. For example, we can have wireless network applications where the large number of users are distributed on a large geographical area. Such networks can potentially face a power limitation since signals loose a lot of power traveling over large distances. However, if the allocated bandwidth is small or the available power per node is large, such networks can still be bandwidth-limited rather than power-limited. On the other hand, one can also imagine wireless network applications where the large number of users are distributed on a relatively small area so that all users are within the transmission range of each other. One expects not to observe any power limitation in such networks. In the wide-band regime however, such a network can still be power limited rather than bandwidth limited. We expect from our discussion on the operating regimes of the AWGN channel that the capacity of the network behaves differently depending on whether the network is power or bandwidth limited. Thus, the right strategies for these different wireless network applications will naturally be different. Moreover, these are the variations we imagine in the light of our understanding of the operating regimes of a point-to-point AWGN channel. Wireless networks comprise a huge number of such point-to-point links. It is also possible that these links have different characteristics, some links may be power limited when some others are bandwidth limited. For example, a wireless network can be locally bandwidth limited while power limited on the global scale. Such a network may exhibit a behavior that can not be anticipated with our understanding of the point-to-point channel. Therefore, one of the main ideas to be conveyed by this dissertation is that the dominant scaling law formulation of wireless networks in the current literature, that focuses on one particular coupling of the area and the number of users in the network, is insufficient in classifying wireless networks. This formulation has created the expectation to describe wireless

networks with a single scaling law formula and to devise a single universal scheme that is optimal for any application. Wireless networks turn out to be far more complicated than this.

In this thesis, we propose to identify system parameters that can have a large range in wireless networks and study all possible interplay between them. These parameters are the area of the network, the bandwidth, the power and the number of users. In complete analogy with the AWGN case, we formulate the interplay as a scaling law problem focusing on the large n limit, but we study now all possible interplays between A, P, W and n. In the most general sense, we let $A = n^{\beta_1}$, $P = n^{\beta_2}$, $W = n^{\beta_3}$ and identify the scaling exponent,

$$e(\alpha, \beta_1, \beta_2, \beta_3) := \lim_{n\to\infty} \frac{\log \rho(\alpha, \beta_1, \beta_2, \beta_3)}{\log n}$$

of the spectral efficiency ρ in bits/s/Hz, for any $\beta_1, \beta_2, \beta_3$. Note that the capacity is given by $C = W\rho$ in bits/s. Indeed in parallel to the AWGN case, the scaling law problem can be expressed in a simpler form. Recall that the transmitted signals are assumed to experience a power path loss of $r^{-\alpha}$ and a random rotation in their phase. For this channel model, it turns out that the spectral efficiency depends on the area of the network, the power and the bandwidth only through a single SNR parameter. In the case of networks there are many SNRs, and one can take any of these different SNRs as reference. Here, without loss of generality, we choose to work with the received SNR over the typical nearest neighbor distance in the network, denoted by SNR_s. Thus, the scaling law problem can be equivalently stated as characterizing the scaling exponent

$$e(\alpha, \beta) := \lim_{n\to\infty} \frac{\log \rho(\alpha, \beta)}{\log n}$$

of the spectral efficiency ρ for any real β where $\mathrm{SNR}_s = n^{\beta}$. In the first and the second chapters of this dissertation, we characterize the scaling exponent $e(\alpha, \beta)$ for any real β and $\alpha \geq 2$. This characterization leads to an approximation of the capacity of large wireless networks, analogous to (1.2) for the AWGN case. The approximation, together with its implications on the operating regimes of large wireless networks, is discussed in the next section.

1.2.1 Operating Regimes of Large Wireless Networks

The scaling law formulation suggested in the previous section allows us to obtain an approximation for the capacity of wireless networks, which identifies four qualitatively different operating regimes, depending on three engineering quantities that stand out in the analysis: the power path loss exponent, the short-distance SNR, and the long-distance SNR. The short-distance SNR is the received SNR in a point-to-point transmission over the typical nearest

neighbor distance inside the network,

$$\mathrm{SNR}_s := \frac{P_r}{N_0 W}, \tag{1.3}$$

where P_r is the received power from a node at the typical nearest neighbor distance. The long-distance SNR is defined as,

$$\mathrm{SNR}_l := n\frac{n^{-\alpha/2}P_r}{N_0 W}, \tag{1.4}$$

where $n^{-\alpha/2}P_r$ is the received power from a node at distance equal to the diameter of the network. Note that in a network where users are uniformly distributed over the two dimensional network area, the typical nearest neighbor distance is $1/\sqrt{n}$ times the diameter of the network. This yields the expression $n^{-\alpha/2}P_r$ for the received power in a point-to-point transmission over a diameter distance. The total capacity C of large wireless networks, in bits/s, exhibits four different behaviors depending on these three parameters:

$$C \propto \begin{cases} nW & \mathrm{SNR}_l \gg 0 \text{ dB} \\ n^{2-\alpha/2}P_r/N_0 & \mathrm{SNR}_l \ll 0 \text{ dB and } 2 \le \alpha \le 3 \\ \sqrt{n}P_r/N_0 & \mathrm{SNR}_s \ll 0 \text{ dB and } \alpha > 3 \\ \sqrt{n}W^{\frac{\alpha-3}{\alpha-2}}(P_r/N_0)^{\frac{1}{\alpha-2}} & \mathrm{SNR}_l \ll 0 \text{ dB, } \mathrm{SNR}_s \gg 0 \text{ dB} \\ & \quad \text{and } \alpha > 3. \end{cases} \tag{1.5}$$

Note two immediate observations. First, there are two SNR parameters of interest in networks, the short and the long distance SNR's, as opposed to the point-to-point case, where there is a single SNR parameter. Second, the most natural way to measure the long-distance SNR in networks is not the SNR of a pair separated by a distance equal to the diameter of the network, but it is n times this quantity as defined earlier in (1.4). There are order n nodes located at a diameter distance to any given node in the network, hence n times the SNR between farthest nodes is the total SNR that can be transferred to this node across this large spatial scale. On the other hand, a node has only a constant number of nearest neighbors, and hence the short-distance SNR in (1.3) is simply the SNR between a nearest neighbor pair. Note that since $\alpha \ge 2$, the long-distance SNR is always smaller than or equal to the short-distance SNR.

The first regime in (1.5) is a degrees of freedom limited regime. The bandwidth and the number of nodes in the network together constitute the available degrees of freedom in the system. In this regime, the network does not face any power limitation, since even the long-distance SNR in the network is large. Thus, long-distance communication is feasible and good communication schemes should exploit this feasibility. On the other hand, the network is degrees of freedom limited, so good communication schemes for this regime should also achieve the full degrees of freedom in the system. These are the properties of the MIMO scheme based on hierarchical cooperation presented

in the previous section: With the help of the hierarchical cooperation architecture, the communication in the network is done via cooperative MIMO transmissions between large clusters of nodes (of size almost of order n) and at distance of the order of diameter of the network. The performance of the MIMO transmissions is linear in the number of nodes, implying that interference limitation is removed by cooperation, and full degrees of freedom are achieved, at least as far as scaling is concerned. The performance in this regime is qualitatively the same as that obtained in the dense scaling.

In all the other regimes, the long-distance received SNR is less than 0 dB. Hence, the network is power-limited and the transfer of power becomes important in determining performance. In the second regime, i.e., when $\alpha \leq 3$, signal power decays slowly with distance and the total power transfer is maximized by global cooperation. Cooperative MIMO communication not only achieves the full degrees of freedom in the system but it also provides a power gain, obtained by combining signals received at different nodes. With the hierarchical cooperation architecture, it allows to combine the received signals by a cluster of nodes, almost the size of the network. This power gain becomes important in this regime and in Chapter 4 of this thesis, we show that a modification of the hierarchical cooperation scheme can achieve optimal capacity scaling. Note that this is a power-limited regime, hence the performance depends critically on the available power, but not so much on the bandwidth.

When $\alpha > 3$, signals decay fast with distance, and the transfer of power is maximized by cooperating at smaller scales. In this case, there is no benefit in combining the signals received by a large cluster of nodes. The total power received by such a large cluster is already dominated by the power received by few nodes in the cluster, located closest to the transmitting nodes. It is more beneficial to perform shorter-range communication between clusters containing fewer nodes. Then, the rest of nodes in the network can undertake simultaneous transmissions, suggesting the idea of spatial reuse. When the nearest-neighbor $\mathrm{SNR}_s \ll 0$ dB (third regime), the communication scale reduces to the nearest neighbor distance. The optimal strategy in this regime is to confine to nearest neighbor transmissions and multi-hop information across the network. The point-to-point nearest-neighbor transmissions are power limited since $\mathrm{SNR}_s \ll 0$ dB, so the overall capacity of the network is also power limited. The extended scaling considered in the literature is qualitatively similar to the second and third regimes.

The most interesting regime and the one that is most counter-intuitive, given our understanding of the point-to-point AWGN channel, is the fourth regime, when $\alpha > 3$ and $\mathrm{SNR}_l \ll 0$ dB, but $\mathrm{SNR}_s \gg 0$ dB. Note that since $\mathrm{SNR}_l \ll 0$ dB, this is still a power limited regime and optimal schemes for this regime should transfer power efficiently across the network. The nearest-neighbor transmissions are now bandwidth-limited and not power-efficient in translating the power gain into capacity gain. There is the potential of increasing throughput by spatially multiplexing transmission via cooperation within

clusters of nodes and performing distributed MIMO. Yet, the clusters cannot be as large as the size of the network, since power attenuates rapidly for $\alpha > 3$. The exact cooperation scale is dictated by the power path loss exponent and the short-distance SNR in the network.

It turns out that the optimal scheme in this regime is to cooperate hierarchically within clusters of an intermediate size, perform MIMO transmission between adjacent clusters and then multi-hop across several clusters to get to the final destination. This multi-hop-MIMO scheme is introduced and analyzed in Chapter 4 of the thesis. The optimal cluster size is chosen such that the received SNR in the MIMO transmission is at 0 dB. Any smaller cluster size results in power inefficiency. Any larger cluster size reduces the amount of spatial reuse without providing any extra benefit in terms of power transfer. The two extremes of this architecture are precisely the traditional multi-hop scheme, where the cluster size is 1 and the number of hops is $\sqrt{n}$, and the long-range cooperative MIMO scheme, where the cluster size is of order n and the number of hops is 1. Note also that because short-range links are bandwidth-limited and long-range links are power-limited, the network capacity is *both* bandwidth and power-limited. Thus, the capacity is sensitive to both the amount of bandwidth and the amount of power available. This regime is fundamentally a consequence of the heterogeneous nature of links in a network and does not occur in point-to-point links, nor in the dense or extended scalings.

1.3 Throughput-Delay Trade-off for Hierarchical Cooperation

In Chapter 5 of the present dissertation, a scaling law formulation is used to study the throughput-delay trade-off for the hierarchical cooperation scheme. We show that a modification of the scheme achieves a throughput-delay tradeoff $D(n) = (\log n)^2 T(n)$ for $T(n)$ between $\Theta(\sqrt{n})$ and $\Theta(n^{1-\epsilon})$ for arbitrarily small $\epsilon > 0$, where $D(n)$ and $T(n)$ are the average delay experienced by bits traveling inside the network and the aggregate throughput, respectively. This result implies that the delay scaling of the hierarchical cooperation scheme is roughly equal to its aggregate throughput scaling. In other words, larger aggregate throughput comes at the expense of larger delay. This trade-off complements the previous results of [21, 22], which show that the throughput-delay trade-off for multi-hopping is given by $D(n) = T(n)$, where $T(n)$ lies between $\Theta(1)$ and $\Theta(\sqrt{n})$.

Besides establishing the throughput-delay trade-off for the hierarchical cooperation scheme, the aim of Chapter 5 is to demonstrate that there is a lot of room for improving the original form of the scheme introduced in Section 4.4. Establishing the throughput-delay tradeoff above requires one such improvement. The presentation in Section 4.4 is focused only on the throughput

performance of the hierarchical cooperation scheme and yields an architecture which is optimal from throughput scaling point of view. The scheme achieves an aggregate throughput $T_h(n) = \Theta(n^{\frac{h}{h+1}})$ for any integer $h > 0$, where h corresponds to the number of hierarchical levels used in the scheme. Increasing h, the scheme gets arbitrarily close to linear scaling. However, the delay of the scheme scales like $D_h(n) = \Theta(n^{\frac{h}{2}})$; in other words, it grows arbitrarily fast as the throughput approaches linear scaling. In Chapter 5, we present a modification of the scheme that achieves the same aggregate throughput $T_h(n) = \Theta(n^{\frac{h}{h+1}})$ with much smaller delay $D_h(n) = \Theta(n^{\frac{h}{h+1}})$.

The key to the modification is a more careful treatment of the cooperation problem for distributed MIMO communication. The cooperation problem is modeled by the following traffic pattern: Assume that each of the nodes in the wireless network wants to communicate an independent message of length L bits to each of the other nodes in the network, for a constant L independent of n. This uniform traffic pattern is different from the permutation traffic that is of main interest in this dissertation. In the case of permutation traffic defined earlier in Section 1.1, each node in the network is source for exactly one communication request and destination for another. Moreover, the interest is in the scaling of the rate of communication, in which case the number of bits communicated between each source-destination pair can increase arbitrarily with the number of users in the network. In the uniform traffic problem considered in Chapter 5, we constrain the number of bits to be communicated between every pair of nodes in the network, to fixed L bits. We are interested in minimizing the total time required to complete this task. We propose a two-phase hierarchical scheme that solves this uniform traffic problem in $\Theta(n^{\frac{h+1}{h}})$ time-slots, for any $h > 0$. This particular result may be of interest in its own right.

2 Model

In this chapter we introduce the model considered in this dissertation and some basic definitions and properties that are commonly used in the following three chapters.

2.1 Model

There are n nodes with transmitting and receiving capabilities that are uniformly and independently distributed in a rectangle of area $\sqrt{A} \times \sqrt{A}$. Each node has an average transmit power budget of P Watts and the network is allocated a total bandwidth of W Hertz around a carrier frequency of f_c, $f_c \gg W$. Every node is both a source and a destination for some traffic request. The sources and destinations are randomly paired up one-to-one into n source-destination pairs without any consideration on node locations. Each source wants to communicate to its destination node at the same rate R in bits/s/Hz. (Note that R corresponds to the spectral efficiency per node as discussed earlier in Section 1.2.) The total throughput of the system is $T = nR$ bits/s/Hz.

We assume that communication takes place over a flat channel and the complex baseband-equivalent channel gain between node i and node k at timeslot m is given by

$$H_{ik}[m] = \sqrt{G} r_{ik}^{-\alpha/2} \exp(j\theta_{ik}[m]) \tag{2.1}$$

where r_{ik} is the distance between the nodes i and k, and $\theta_{ik}[m]$ is the random phase at time m, uniformly distributed in $[0, 2\pi]$. We assume that $\{\theta_{ik}[\cdot],\ 1 \leq i, k \leq n,\ i \neq k\}$ is a collection of independent identically distributed random processes that are also independent of the locations r_{ik}, $1 \leq i, k \leq n$, $i \neq k$. Note that the channel is random, depending on the location of the users

and the phases. The locations are assumed to be fixed over the duration of the communication. The phases are assumed to vary in a stationary ergodic manner (fast fading). We assume that the channel gains are known at all the nodes.

The parameters G and $\alpha \geq 2$ are assumed to be constants; α is called the power path loss exponent. For example, under free-space line-of-sight propagation, Friis' formula applies and

$$|H_{ik}[m]|^2 = \frac{G_{Tx} \cdot G_{Rx}}{(4\pi r_{ik}/\lambda_c)^2} \tag{2.2}$$

so that

$$G = \frac{G_{Tx} \cdot G_{Rx} \cdot \lambda_c^2}{16\pi^2}, \qquad \alpha = 2$$

where G_{Tx} and G_{Rx} are the transmitter and receiver antenna gains respectively and λ_c is the carrier wavelength. The discrete-time complex baseband signal received by node i at time m is given by

$$Y_i[m] = \sum_{\substack{k=1 \\ k \neq i}}^{n} H_{ik}[m]X_k[m] + Z_i[m] \tag{2.3}$$

where $X_k[m]$ is the signal sent by node k at time m subject to an average power constraint

$$\mathbb{E}(|X_k|^2) \leq P/W$$

and $Z_i[m]$ is complex white circularly symmetric Gaussian noise of variance N_0.

The channel model we introduce above is a well-established one, used in almost all works in wireless communication and all the earlier works on scaling laws. Several comments about this model are in order:

- The path loss model is based on a *far-field* assumption: the distance r_{ik} is assumed to be much larger than the carrier wavelength. This is typically the case in wireless network applications. When the distance is of the order or shorter than the carrier wavelength, the simple path loss model obviously does not hold anymore as path loss can potentially become path "gain". The reason is that near-field electromagnetics now come into play.

- The phase $\theta_{ik}[m]$ depends on the distance between the nodes modulo the carrier wavelength. The random phase model is thus also based on a far-field assumption: we assume the nodes' separation is at a much larger spatial scale compared to the carrier wavelength, so that the phases can be modeled as completely random and independent of the actual positions. When the nodes are packed close together, the independence assumption of the phases breaks down. This case has been addressed

in the work [42], which has been conducted as a part of this thesis, but the results are not included in the current write-up. By considering the extremal model with no random phases, or equivalently when all the phases are fully correlated and equal to each other, we show in [42] that the conclusions can differ significantly from those presented in this dissertation based on the i.i.d. random phase channel model. A more refined result has been reported in a recent work [19], which roughly implies that the conclusions of the i.i.d. phase model hold as long as the separation between nodes is larger than $\lambda_c\sqrt{n}$. For a carrier frequency of 3 GHz, corresponding to a carrier wavelength of 0.1 m, the result of [19] implies that the average separation between users should be at least 3 m in a network of 1000 nodes and 10 m in a network of 10000.

- It is realistic to assume the time-variation of the phases, since they vary significantly when users move a distance of the order of the carrier wavelength (fractions of a meter). The positions determine the path losses and they, on the other hand, vary over a much larger spatial scale. So the positions are assumed to be fixed. The results of this dissertation can also be extended to the slow fading setting where the phases are drawn from an independent uniform distribution and are kept fixed during communication. The slow fading assumption allows for a simpler derivation of the results in Chapter 3 and requires an extra technical step to extend the results of Chapters 4 and 5. These extensions are not included in the current dissertation. The key insight behind the extension is the self-averaging effect of a large number of independent random variables.

- The random phase model essentially corresponds to a line-of-sight type environment and ignores multipath effects. The randomness in phases is sufficient to achieve full degrees of freedom in the network. With multipaths, there is a further randomness due to random constructive and destructive interference of these paths. The extension of the results in this dissertation to the multipath case is straightforward.

2.2 Definitions

Given the network model in Section 2, let us define the received SNR in a point-to-point transmission over the typical nearest neighbor distance to be,

$$\mathrm{SNR}_s := \frac{GP}{N_0 W(\sqrt{A/n})^\alpha}. \tag{2.4}$$

where $\sqrt{A/n}$ is the typical nearest neighbor distance in the network. This quantity will be often referred to as the short-distance SNR in the network. Let us analogously define the long-distance SNR in the network as

$$\mathrm{SNR}_l := n\frac{GP}{N_0\, W\, (\sqrt{A})^\alpha}. \tag{2.5}$$

The conclusions derived in this dissertation on the capacity scaling and optimal operation of wireless networks will depend on these two SNR parameters. However, for the derivations in the next two chapters, it will be also useful to define a more general version of these quantities, namely the received SNR over any spatial scale $\sqrt{A_c}$, where $\sqrt{A_c}$ can range from the nearest neighbor distance $\sqrt{A/n}$ to the diameter of the network $\sqrt{A}$. We define

$$\mathrm{SNR}(A_c) := \frac{A_c n}{A} \frac{GP}{N_0 W (\sqrt{A_c})^\alpha}, \qquad \text{for} \quad \frac{A}{n} \leq A_c \leq A. \tag{2.6}$$

Note that $\frac{GP}{N_0 W (\sqrt{A_c})^\alpha}$ is the received SNR in a point-to-point transmission over the distance $\sqrt{A_c}$. In (2.6), we multiply this quantity by $\frac{A_c n}{A}$, the number of nodes typically contained in a cluster of area A_c. Thus, the quantity $\mathrm{SNR}(A_c)$ can be interpreted as the total SNR that can be transferred to a node over the spatial scale $\sqrt{A_c}$ since $\frac{A_c n}{A}$ is an order estimate of the total number of nodes in the network located at distance $\sqrt{A_c}$ to a given node. Note that the short-distance SNR and the long distance SNR are the two extremes of $\mathrm{SNR}(A_c)$: when $\sqrt{A_c}$ is the minimal possible spatial scale in the network, the nearest-neighbor distance, $\mathrm{SNR}_s = \mathrm{SNR}(A/n)$ and when $\sqrt{A_c}$ is the largest possible scale in the network, the diameter, $\mathrm{SNR}_l = \mathrm{SNR}(A)$. Note also that in a network with uniformly distributed nodes, the short-distance SNR and the long-distance SNR are related by the equation,

$$\mathrm{SNR}_l = n^{1-\alpha/2} \mathrm{SNR}_s. \tag{2.7}$$

Note that since $\alpha \geq 2$, $\mathrm{SNR}(A_c)$ is non-increasing in its argument A_c. When $\alpha = 2$, $\mathrm{SNR}_s = \mathrm{SNR}(A_c) = \mathrm{SNR}_l$ for any $\frac{A}{n} \leq A_c \leq A$. Otherwise for $\alpha > 2$, $\mathrm{SNR}(A_c)$ is strictly decreasing in A_c.

The interest in this thesis is to devise schemes that achieve optimal throughput scaling in wireless networks. A scheme and its throughput are defined as follows.

Definition 2.2.1 (Scheme)**.** *A scheme for the random network model described in the previous section is a sequence Π_n of communication techniques and coordinated network protocols, where Π_n describes how communication between the n source-destination pairs takes place in any random realization of the network with n nodes.*

Definition 2.2.2 (Throughput of a Scheme)**.** *A scheme Π_n is said to achieve an aggregate throughput T bits/s/Hz with high probability, if in a given realization of the network each of the n source-destination pairs communicates at least $B(t)$ bits in t time slots under this scheme and*

$$\lim_{n\to\infty} \mathbb{P}\left(\lim\inf_{t\to\infty} \frac{1}{t} B(t) \geq \frac{T}{n} \right) = 1.$$

where the probability is over the random realizations of the network, i.e., the random distribution of nodes over the network area and the random source-destination pairings.

In this dissertation, we restrict attention to the scaling of the throughput achieved by a particular scheme with the number of users in the network. The scaling exponent of the total throughput T is defined as,

$$e(\alpha, \beta) := \lim_{n\to\infty} \frac{\log T}{\log n}. \tag{2.8}$$

where

$$\beta := \lim_{n\to\infty} \frac{\log \mathrm{SNR}_s}{\log n}. \tag{2.9}$$

is the scaling exponent of SNR_s. We assume that β is a finite real number. Note that the scaling of SNR_s is jointly dictated by the scalings of system parameters P, A and W with n.

To describe limiting behavior of functions, we often adopt the following notation throughout this dissertation: For two functions $f(n)$ and $g(n)$, the notation $f(n) = O\left(g(n)\right)$ means that $|f(n)/g(n)|$ remains bounded as n increases. We write $g(n) = \Theta\left(f(n)\right)$ to denote that $f(n) = O\left(g(n)\right)$ and $g(n) = O\left(f(n)\right)$. Finally, $f(n) = \Omega\left(g(n)\right)$ if $|g(n)/f(n)|$ remains bounded as n increases.

2.3 Properties of the Random Network

In the following lemma we state several properties that are satisfied with high probability in a random realization of the network. For a sequence of random variables A_n and a sequence of numbers b_n,

$$A_n \leq b_n, \qquad \text{with high probability (w.h.p.)}$$

if

$$\lim_{n\to\infty} \mathbb{P}(A_n \leq b_n) = 1.$$

The regularity properties given below arise from the assumption that nodes are distributed uniformly at random over the network area and source-destination pairings are also formed randomly without any consideration of node locations. These properties will be used repeatedly in the following chapters. The proof of the lemma is given in Appendix 2.A.

Lemma 2.3.1. *The random network of n nodes with area A and random source-destination pairings satisfies the following properties:*

a) Consider a cut dividing the network area into two equal halves. The number of source-destination pairs with sources on the left-half network and destinations on the right-half network is in the interval $((1-\delta)n/4, (1+\delta)n/4)$, for any $\delta > 0$, w.h.p.

b) The minimal distance between any two nodes in the network is larger than $\sqrt{A}/n^{1+\delta}$, for any $\delta > 0$, w.h.p.

c) *Let the network area be divided into n cells of area A/n. Then, there are less than $\log n$ nodes inside each cell, w.h.p.*

d) *Let the network area be divided into $n/2\log n$ cells each of area $2A\log n/n$. Then, there is at least one node inside each cell, w.h.p.*

e) *Let us partition the network area A into cells of area A_c, where A_c can be a function of n and A. For any $0<\delta<1$, the number of nodes inside each cell is in the interval $\left((1-\delta)\frac{A_c}{A}n,(1+\delta)\frac{A_c}{A}n\right)$ with probability larger than $1-\frac{2A}{A_c}e^{-\Lambda(\delta)\frac{A_c}{A}n}$, where $\Lambda(\delta)$ is independent of n, A and A_c, and satisfies $\Lambda(\delta)>0$ when $\delta>0$.*

2.A Regularity Properties of Random Networks

In this section, we prove some regularity properties satisfied with high probability in a random realization of the network. Recall that we have n nodes, paired up into n source-destination pairs. These n nodes are independently distributed on the network area A, such that the position of each node is a uniform random variable over the network area.

Proof of Lemma 2.3.1: Parts-(a) and (c) of the lemma are proven after part-(e).

b) Consider one specific node in the network which is at distance larger than $\sqrt{A}/n^{1+\delta}$ to all other nodes in the network for some $\delta>0$. This is equivalent to saying that there are no other nodes inside a circle of area $\pi A\,/\,n^{2+2\delta}$ around this node. The probability of such an event is

$$\left(1-\frac{\pi}{n^{2+2\delta}}\right)^{n-1}.$$

Moreover, the minimum distance between any two nodes in the network is larger than $\sqrt{A}/n^{1+\delta}$ if and only if this condition is satisfied for all nodes in the network. Thus by the union bound we have,

$$P\left(\text{minimum distance in the network is smaller than }\frac{\sqrt{A}}{n^{1+\delta}}\right)$$
$$\leq n\left(1-\left(1-\frac{\pi}{n^{2+2\delta}}\right)^{n-1}\right)$$

which decreases to zero as $1/n^{2\delta}$ with increasing n. Therefore, the minimal distance between any two nodes in the network is larger than $\sqrt{A}/n^{1+\delta}$, for any $\delta>0$ w.h.p.

d) The probability that a given cell is empty is given by $\left(1-\frac{\log n}{n}\right)^n$. The probability that there exists an empty cell is bounded above by

$$n\left(1-\frac{2\log n}{n}\right)^n,$$

hence decreases to zero as $1/n$ when n is large.

e) The proof of the statement is a standard application of the exponential Chebyshev's inequality. Note that the number of nodes in a given cell is a sum of n i.i.d Bernoulli random variables B_i, such that $\mathbb{P}(B_i = 1) = \frac{A_c}{A}$. For any $s > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} B_i \geq (1+\delta)\frac{A_c}{A}n\right) = \mathbb{P}\left(e^{s\sum_{i=1}^{n} B_i} \geq e^{s(1+\delta)\frac{A_c}{A}n}\right) \tag{2.10}$$

$$\begin{aligned} &\leq \left(\mathbb{E}[e^{sB_1}]\right)^n e^{-s(1+\delta)\frac{A_c}{A}n} \\ &= \left(e^s\frac{A_c}{A} + (1-\frac{A_c}{A})\right)^n e^{-s(1+\delta)\frac{A_c}{A}n} \\ &\leq e^{\frac{A_c}{A}n(e^s-1)}\, e^{-s(1+\delta)\frac{A_c}{A}n} \\ &= e^{-\frac{A_c}{A}n\Lambda_+(\delta)} \end{aligned} \tag{2.11}$$

by choosing $s = \ln(1+\delta)$, where $\Lambda_+(\delta) = (1+\delta)\ln(1+\delta) - \delta$. Note that $\Lambda_+(\delta) > 0$ when $\delta > 0$. The probability of having a cell with more than $(1+\delta)\frac{A_c}{A}n$ nodes is upperbounded by the union bound as

$$\mathbb{P}\left(\exists \text{ a cell with } \#\text{ of nodes } \geq (1+\delta)\frac{A_c}{A}n\right) \leq \frac{A}{A_c}e^{-\frac{A_c}{A}n\Lambda_+(\delta)}.$$

The proof for the lower bound follows similarly and yields

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{n} B_i \leq (1-\delta)\frac{A_c}{A}n\right) &= \mathbb{P}\left(e^{-s\sum_{i=1}^{n} B_i} \geq e^{-s(1-\delta)\frac{A_c}{A}n}\right) \\ &\leq e^{-\frac{A_c}{A}n\Lambda_-(\delta)} \end{aligned}$$

by choosing $s = -\ln(1-\delta)$, where $\Lambda_-(\delta) = (1-\delta)\ln(1-\delta) + \delta$. The conclusion follows by defining $\Lambda(\delta) = \min(\Lambda_-(\delta), \Lambda_+(\delta))$.

a) A straightforward application of part-(c) shows that the number of nodes contained in the left-half network, n_l, is such that $(1-\delta_1)n/2 \leq n_l \leq (1+\delta_1)n/2$ w.h.p for any $0 < \delta_1 < 1$. All these nodes are sources for some traffic. Let n_{lr} of the n_l destination nodes corresponding to these n_l source nodes be located in the right half network. A second application of part-(c) yields $(1-\delta_2)n_l/2 \leq n_{lr} \leq (1+\delta_2)n_l/2$ w.h.p. for any $0 < \delta_2 < 1$. The result follows by combining the two bounds, yielding $(1-\delta_3)n/4 \leq n_{lr} \leq (1+\delta_3)n/4$ w.h.p. for any $0 < \delta_3 < 1$.

c) The statement can be proved by following the upperbound derivation in (2.11). The probability that a given cell of area A/n contains more than $\log n$ nodes is given by

$$\mathbb{P}\left(\sum_{i=1}^{n} B_i \geq \log n\right) \leq e^{e^s-1}\, e^{-s\log n},$$

for any $s > 0$. Choosing $s > 2$ and applying the union bound, it can be shown that the probability that there exist a cell containing more than $\log n$ nodes decreases to zero as $1/n$ when n increases.

3 Upper Bound

In this chapter, we derive an information theoretic upper bound on the capacity scaling of wireless networks. The upper bound is information theoretic because it emerges from basic assumptions on the physical channel and the network, and no specific assumption is made about the communication or networking technique employed. As such, it characterizes the ultimate limit of performance in wireless networks and applies globally to any possible network communication scheme. We will see in the next chapter that the upper bound is indeed tight, as there are communication schemes that can achieve this performance. Thus, the current and the next chapter together characterize the capacity scaling of wireless networks for the model described in Chapter 2. This model is a generalization of the model initially suggested by [27]. Section 3.1 presents an overview of the results of [27], as well as the follow-up work in the literature on upper bounds for wireless networks. Section 3.1 contains the main derivation of the upper bound. The derivation reveals four qualitatively different operating regimes in wireless networks. In the last section of this chapter, we discuss the insights suggested by the upper bound derivation on the properties of optimal schemes for each regime. These optimal schemes are constructed in the next chapter.

3.1 Existing Upper Bounds on the Capacity of Wireless Networks

The random network model presented in the previous section has been first proposed by [27], though in a more restricted form. The restriction is that independent parameters such as the area of the network, the number of nodes contained in the network, the power and bandwidth budget are coupled to-

gether in a specific way: The area $2A$, the power P and the bandwidth W remain constant as the number of nodes n increases in the network. This particular way of scaling the parameters of the network has been widely referred to as *dense scaling* or *dense network* later on. The work [27] derives a "semi-information theoretic" upper bound on the capacity scaling of such dense networks. The upper bound is "semi-information theoretic", as all communication in the network is a priori restricted to be of the form of point-to-point communication: the signals received from nodes other than one particular transmitter are to be regarded as noise degrading the communication link. As such, the upper bound of [27] shows that classical network architectures with conventional single-user decoding and forwarding of packets cannot achieve a total throughput scaling better than $O(\sqrt{n})$. Equivalently, the rate of communication between any source-destination pair in the network has to decrease to zero as $O(\frac{1}{\sqrt{n}})$, when the number of users n becomes large. The intuition behind this upper bound is the following: Given the restriction to point-to-point communication, the only way to deliver packets to their destinations is via multi-hopping and essentially, the only flexibility left in the design is the range of these hops. It turns out that long-range communication is not beneficial, as the interference generated would preclude most of the other nodes from communicating. Instead, it is better to stick to nearest neighbor communication and maximize the number of simultaneous transmissions (spatial reuse). However, this means that each packet has to be retransmitted many times before getting to its final destination, leading to a sub-linear scaling of system throughput.

The work [27] gives a very good insight on the potential, as well as the limitations of the current physical-layer technology in wireless networks. However, from an information theoretic perspective, this is a very restricted treatment of wireless networks. The inherent broadcasting and superposition nature of wireless communication, as well as the fact that there are many nodes involved, each with transmitting and receiving capabilities, makes this setup very rich in terms of possibilities for enhancing information transfer. Hence, one wishes to study wireless networks without making assumptions rooted in the current engineering practice, since such assumptions may not be applicable in general and thus essentially are arbitrary. This raises the question of whether the $O(\sqrt{n})$ upper bound of [27] is indeed universal and applies to any possible scheme that does not necessarily confine to point-to-point communication. This question was first addressed in [54] where it was shown that whenever the power path loss exponent α of the environment in (2.1) is greater than 6 (i.e. the received power decays faster than r^{-6} with the distance r from the transmitter), $O(\sqrt{n})$ upper bounds the total throughput scaling of any possible scheme. The work [54] was followed by several others [30, 34, 53, 3]. Successively, they improved the threshold on the path loss exponent α for which the $O(\sqrt{n})$ scaling law of [27] could be confirmed information theoretically ($\alpha > 5$ in [30], $\alpha > 4.5$ in [3] and $\alpha > 4$ in [53]). The question has been open for the important range of α between 2 and 4. A corollary of the results presented in the next section of

this thesis fills this gap.

However, there is a subtle point that has received little attention in the literature up-to now. Although [54] and the follow-up works set to confirm the "semi-information theoretic" upper bound of [27] from an information theoretic point of view, they are based on different network models. The paper [27] deals with dense networks while [54] and the subsequent works couple the network parameters in a different way. The area $2A = n$, while the power P and the bandwidth W remain constant as the number of nodes n increases in the network. This way of scaling the parameters of the network is referred to as *extended scaling* or *extended network*. The results of the next section will reveal that the dense and the extended scalings are indeed fundamentally different from each other. A way to understand the difference between the engineering implications of these two network scalings is by drawing a parallel with the classical notions of *interference-limitedness* and *coverage-limitedness*, the two operating regimes of cellular networks. Cellular networks in urban areas tend to have dense deployments of base-stations, so that signals are received at the mobiles with sufficient signal-to-noise ratio (SNR), but performance is limited by *interference* between transmissions in adjacent cells. Cellular networks in rural areas, on the other hand, tend to have sparse deployments of base-stations, so that performance is mainly limited by the ability to transmit enough power to reach all the users with sufficient signal-to-noise ratio. Analogously, in the dense network scaling, all nodes can communicate with each other with sufficient SNR; performance can only be limited by interference, if at all. The $O(\sqrt{n})$ upper bound of Gupta-Kumar comes precisely from such an interference limitation. In the extended network scaling, the source and destination pairs are at increasing distance from each other, and so both interference limitation and power limitation can come into play. The network can be coverage-limited and/or interference-limited. The information-theoretic limit on performance proved in [54, 30, 34, 53, 3] are all based on the cutset bound [12, Theorem 14.10.1], that assumes full cooperation between the transmitting and receiving nodes. Hence, starting with a cooperative bound, the limitation captured by these works is not due to interference. The limitation comes from bounding the maximum amount of power that can be transferred across the network. What was shown by these works is that for $\alpha > 4$, when signals attenuate fast enough, the extended network is fundamentally coverage-limited: even with optimal cooperative relaying, the amount of power transferred across the network does not allow to achieve a throughput scaling better than $O(\sqrt{n})$. In this sense, these works have not answered the original question of whether the interference limitation implied by the "semi-information theoretic" upper bound of [27] is fundamental or not, but they have introduced a new phenomenon, power limitation in wireless networks, into the discussion. In the following section, we will see that the extended scaling is not only power limited for $\alpha > 4$, but for all $\alpha > 2$. Though, the $O(\sqrt{n})$ upper bound holds only for $\alpha \geq 3$. When $2 \leq \alpha < 3$, the total capacity scaling is upper bounded by $O(n^{2-\alpha/2})$. The key to our result on extended scaling is a careful evaluation

of the maximum amount of power that can be transferred across the extended network, by using techniques from random matrix theory. Earlier works [54]–[3] are also based on bounding the maximum power transfer, but their upper bounding techniques are loose for $2 \leq \alpha < 4$, i.e., when attenuation is lower and power transfer becomes easier.

However, the result presented in this chapter is much more general than a tight upper bound on the extended scaling. We present a tight upper bound on the capacity scaling of wireless networks with area $A = n^\gamma$, *for any* γ. It turns out that the problem is indeed much more interesting for $0 < \gamma < 1$, than it is for extended networks (corresponding to $\gamma = 1$) or dense networks (corresponding to $\gamma = 0$). We will see that extended networks correspond to the special case of $\mathrm{SNR}_s = 0$ dB, inside a larger class of networks with $\mathrm{SNR}_s \leq 0$ dB. Such networks are completely power-limited, since even the nearest-neighbor links are in the power-limited regime. On the other hand, dense networks are one special case of networks that experience no power-limitation, i.e., $\mathrm{SNR}_l \geq 0$ dB. Naturally, the most interesting case is when the the network is only partially power-limited, i.e., when $\mathrm{SNR}_s > 0$ dB but $\mathrm{SNR}_l < 0$ dB. This models the common scenario where the channels between different node pairs are in different SNR regimes, short scale links are bandwidth-limited while long-range links are power-limited. Such networks exhibit an interesting behavior that can not be anticipated by studying the dense or extended scalings.

The main result of this chapter is summarized in the next section.

3.2 Main Result

Recall the definition of the scaling exponent of the total throughput T defined earlier in Chapter 2,

$$e(\alpha, \beta) = \lim_{n\to\infty} \frac{\log T}{\log n}.$$

where

$$\beta = \lim_{n\to\infty} \frac{\log \mathrm{SNR}_s}{\log n}.$$

is the scaling exponent of

$$\mathrm{SNR}_s = \frac{GP}{N_0 W (\sqrt{A/n})^\alpha}. \tag{3.1}$$

The main result of this chapter is to establish the following tight upper bound on the aggregate throughput scaling achieved by any scheme in the network. The following section is devoted to the proof of this theorem.

Theorem 3.2.1. *The scaling exponent of the aggregate throughput T is bounded*

above with high probability by,

$$e(\alpha,\beta) \leq \begin{cases} 1 & \beta \geq \alpha/2 - 1 \\ 2 - \alpha/2 + \beta & \beta < \alpha/2 - 1 \text{ and } 2 \leq \alpha < 3 \\ 1/2 + \beta & \beta \leq 0 \text{ and } \alpha \geq 3 \\ 1/2 + \beta/(\alpha - 2) & 0 < \beta < \alpha/2 - 1 \text{ and } \alpha \geq 3, \end{cases} \tag{3.2}$$

for $\alpha \geq 2$ and any real β where β is the scaling exponent of the nearest neighbor SNR.

The upper bounds for the dense and extended scalings can be found as special cases of the above result. In the dense scaling, the area, the bandwidth and the power are constants that do not depend on n. It can be observed from (3.1) that $\mathrm{SNR}_s = \Theta(n^{\alpha/2})$, or equivalently dense networks correspond to $\beta = \alpha/2$ which falls in the first regime in (3.2) yielding an exponent $e(\alpha, \alpha/2) \leq 1$. In the extended scaling $A = n$ while P and W are constants independent of n. In (3.1), $\mathrm{SNR}_s = \Theta(1)$ or equivalently $\beta = 0$. Thus depending on the power path loss exponent, extended networks fall in either the second or the third regime in (3.2), with an exponent equal to

$$e(\alpha, 0) \leq \begin{cases} 2 - \alpha/2 & 2 \leq \alpha \leq 3 \\ 1/2 & \alpha > 3, \end{cases} .$$

Note that neither the dense nor the extended scaling touches the fourth regime.

3.3 Cutset Upper Bound

We consider a cut dividing the network area into two equal halves. We are interested in bounding above the sum of the rates of communications passing through the cut from left to right. These communications with source nodes located on the left and destination nodes located on the right half domain are depicted in bold lines in Fig. 3.1. By Part-(d) of Lemma 2.3.1, this sum-rate is equal to 1/4'th of the total throughput T with high probability. The maximum achievable sum-rate between these source-destination pairs is bounded above by the capacity of the MIMO channel between all nodes S located to the left of the cut and all nodes D located to the right. Under the fast fading assumption, we have

$$T_{L\to R} \leq \max_{\substack{Q(H)\geq 0 \\ \mathbb{E}(Q_{kk}(H))\leq P/W, \forall k\in S}} \mathbb{E}\left(\log\det\left(I + \frac{1}{N_0} HQ(H)H^*\right)\right) \tag{3.3}$$

where

$$H_{ik} = \frac{\sqrt{G}\, e^{j\,\theta_{ik}}}{r_{ik}^{\alpha/2}}, \quad k \in S, i \in D.$$

The mapping $Q(\cdot)$ is from the set of possible channel realizations H to the set of positive semi-definite transmit covariance matrices. The diagonal element

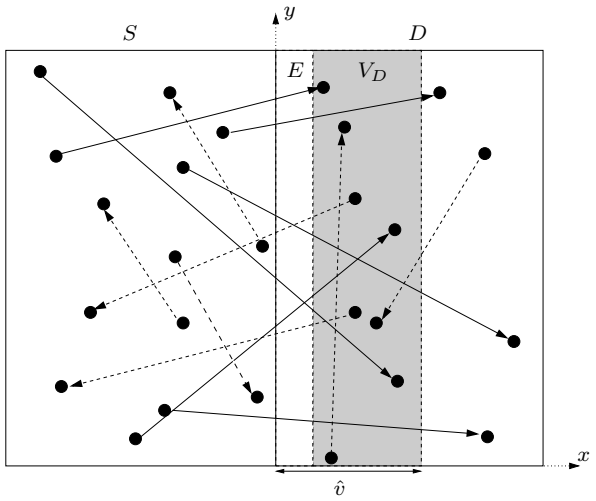

Figure 3.1: The cut-set considered in Section 3.3. The communication requests that pass across the cut from left to right are depicted in bold lines.

$Q_{kk}(H)$ corresponds to the power allocated to the kth node for channel state H. Let us simplify notation by introducing the nearest neighbor SNR defined earlier in (2.4) and also rescale the distances in the network by this nearest neighbor distance, defining

$$\hat{r}_{ik} := \frac{1}{\sqrt{A/n}}\, r_{ik} \quad \text{and} \quad \hat{H}_{ik} := \frac{e^{j\,\theta_{ik}}}{\hat{r}_{ik}^{\alpha/2}}. \tag{3.4}$$

Note that the first transformation rescales space and maps our original network of area $\sqrt{A} \times \sqrt{A}$ to a network of area $\sqrt{n} \times \sqrt{n}$. Consequently, the matrix $\hat{H}$ defined in terms of the rescaled distances relates to such a network with area n. Normalizing the typical nearest neighbor distance to 1 provides the convenience that the received SNR in a point-to-point transmission between two nodes at rescaled distance $\hat{r}$ can be simply written as $\mathrm{SNR}_s \hat{r}^{-\alpha}$. We can thus rewrite (3.3) in terms of these new variables as[1]

$$T_{L\to R} \leq \max_{\substack{Q(\hat{H})\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}))\leq 1,\, \forall k\in S}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s\, \hat{H}Q(\hat{H})\hat{H}^*)\right). \tag{3.5}$$

One way to upper bound (3.5) is through upper bounding the capacity by the total received SNR, formally using the relation

$$\log\det(I + \mathrm{SNR}_s\, \hat{H}Q(\hat{H})\hat{H}^*) \leq \mathrm{Tr}\left(\mathrm{SNR}_s\, \hat{H}Q(\hat{H})\hat{H}^*\right). \tag{3.6}$$

[1] Networks with area extending linearly in the number of users are usually called extended networks in the literature. By rescaling distances we map our original network to such an extended network. However, the problem itself does not reduce to the extended scaling problem, since here we do not necessarily assume $\mathrm{SNR}_s = 1$. Indeed, we keep full generality and are interested in characterizing the whole regime $\mathrm{SNR}_s = n^\beta$, where β can be any real number.

The upper bound is tight only if the SNR received by each right-hand side node, i.e, each diagonal entry of the matrix $\mathrm{SNR}_s\, \hat{H}Q(\hat{H})\hat{H}^*$, is small. (Note that the relation in (3.6) relies on the inequality $\log(1+x) \leq x$, which is only tight if x is small.) Whether this is the case or not depends on SNR_s. It can be shown that if $\mathrm{SNR}_s \leq 1$, the network is highly power-limited and the received SNR is small, i.e., decays to zero with increasing n, for every right-hand side node. Using (3.6) will yield a tight upper bound in that case. However, in the general case SNR_s can be arbitrarily large, which can result in high received SNR for certain right-hand side nodes that are located close to the cut or even for all nodes in D. Hence, before using (3.6), we need to distinguish between those nodes in D that receive high SNR and those that have poor power connections to the left-hand side.

Assumption 3.3.1. *For the sake of simplicity in presentation, we assume in this section that there is a rectangular region located immediately to the right of the cut that is cleared of nodes. Formally, we assume that the set of nodes* $E = \{i \in D : 0 \leq \hat{x}_i \leq 1\}$ *is empty, where* $\hat{x}_i$ *denotes the horizontal coordinate of the rescaled position* $\hat{r}_i = (\hat{x}_i, \hat{y}_i)$ *of node* i*. In fact, w.h.p this property does not hold in a random realization of the network. However, making this assumption allows to exhibit the central ideas in the following derivation in a simpler manner. The extension of the analysis to the general case (without this particular assumption) is given in Appendix 3.B.*

Let V_D denote the set of nodes located on a rectangular strip immediately to the right of the empty region E. Formally, $V_D = \{i \in D : 1 \leq \hat{x}_i \leq \hat{v}\}$, where $1 \leq \hat{v} \leq \sqrt{n}/2$ and $\hat{v} - 1$ is the rescaled width of the rectangular strip V_D. See Fig. 3.1. We would like to tune $\hat{v}$ so that V_D contains the right-hand side nodes with high received SNR from the left-hand side; i.e., those nodes that receive SNR larger than a threshold, say 1. Note however that we do not yet know the covariance matrix Q of the transmissions from the left-hand side nodes, which is to be determined from the maximization problem in (3.5). Thus, we cannot really compute the received SNR of a right-hand side node. For the purpose of specifying V_D however, let us arbitrarily look at the case when Q is the identity matrix and define the received SNR of a right-hand side node $i \in D$ when left-hand side nodes are transmitting *independent* signals at full power to be

$$\mathrm{SNR}_i := \sum_{k \in S} \frac{P}{N_0 W} |H_{ik}|^2 \;=\; \mathrm{SNR}_s \sum_{k \in S} |\hat{H}_{ik}|^2 = \mathrm{SNR}_s\, \hat{d}_i. \tag{3.7}$$

where we have defined

$$\hat{d}_i := \sum_{k \in S} |\hat{H}_{ik}|^2. \tag{3.8}$$

Later, we will see that this arbitrary choice of identity covariance matrix is indeed a reasonable one (Lemma 3.3.1). A good approximation for $\hat{d}_i$ is

$$\hat{d}_i \approx \hat{x}_i^{2-\alpha} \tag{3.9}$$

where $\hat{x}_i$ denotes the rescaled horizontal coordinate of the right-hand side node $i \in D$. (This fact is made precise in Lemma 3.3.3.) Recall that $1 \leq \hat{x}_i \leq \sqrt{n}/2$ and since $\alpha \geq 2$, $\hat{d}_i$ is decreasing in $\hat{x}_i$. Using (3.7) and (3.9), we can identify three different regimes and specify $\hat{v}$ accordingly:

1) If $\mathrm{SNR}_s \geq n^{\alpha/2-1}$, then $\mathrm{SNR}_i \gtrsim 1, \forall i \in D$. Hence, let us choose $\hat{v} = \sqrt{n}/2$ or equivalently $V_D = D$ in this case.

2) If $\mathrm{SNR}_s \leq 1$, then $\mathrm{SNR}_i \lesssim 1, \forall i \in D$. Thus, let us choose $\hat{v} = 1$ or equivalently $V_D = \emptyset$.[2]

3) If $1 < \mathrm{SNR}_s < n^{\alpha/2-1}$, then let us choose

$$\hat{v} = \begin{cases} \sqrt{n}/2 & \text{if} \quad \alpha = 2 \\ \mathrm{SNR}_s^{\frac{1}{\alpha-2}} & \text{if} \quad \alpha > 2 \end{cases}$$

so that we ensure $\mathrm{SNR}_i \gtrsim 1, \forall i \in V_D$.

We now would like to break the information transfer from the left-half domain S to the right-half domain D in (3.5) into two terms. The first term governs the information transfer from S to V_D. The second term governs the information transfer from S to the remaining nodes on the right-half domain, i.e., $D \setminus V_D$. Recall that the characteristic of the nodes V_D is that they have good power connections to the left-hand side, that is, the information transfer from S to V_D is not limited in power, but can be limited in degrees of freedom. Thus, it is reasonable to bound the rate of this first information transfer by the cardinality of the set V_D, rather than the total received SNR. On the other hand, the remaining nodes in $D \setminus V_D$ have poor power connections to the left-half domain and the information transfer to these nodes is limited in power, hence using (3.6) is tight. Formally, we proceed by applying the generalized block Hadamard's inequality (also known as Fischer's inequality) which yields

$$\begin{aligned} \log\det(I + \mathrm{SNR}_s \hat{H} Q(\hat{H}) \hat{H}^*) \leq \log\det(I + \mathrm{SNR}_s \hat{H}_1 Q(\hat{H}) \hat{H}_1^*) \\ + \log\det(I + \mathrm{SNR}_s \hat{H}_2 Q(\hat{H}) \hat{H}_2^*) \end{aligned}$$

where $\hat{H}_1$ and $\hat{H}_2$ are obtained by partitioning the original matrix $\hat{H}$: $\hat{H}_1$ is the rectangular matrix with entries $\hat{H}_{ik}, k \in S, i \in V_D$ and $\hat{H}_2$ is the rectangular matrix with entries $\hat{H}_{ik}, k \in S, i \in D \setminus V_D$. In turn, (3.5) is bounded above by

$$\begin{aligned} T_{L\to R} \leq & \max_{\substack{Q(\hat{H}_1)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_1))\leq 1, \forall k \in S}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s \hat{H}_1 Q(\hat{H}_1) \hat{H}_1^*)\right) \\ & + \max_{\substack{Q(\hat{H}_2)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_2))\leq 1, \forall k \in S}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s \hat{H}_2 Q(\hat{H}_2) \hat{H}_2^*)\right) \end{aligned} \tag{3.10}$$

[2]Note that here we make use of the earlier assumption of an empty strip E of width 1. Without the assumption, we would need to choose $\hat{v} < 1$ in this part.

The first term in (3.10) can be bounded by applying Hadamard's inequality once more, or equivalently, by considering the sum of the capacities of the individual multiple-input single-output (MISO) channels between nodes in S and each node in V_D,

$$\max_{\substack{Q(\hat{H}_1)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_1))\leq 1,\,\forall k\in S}} \mathbb{E}\left(\log\det(I+\mathrm{SNR}_s\hat{H}_1Q(\hat{H}_1)\hat{H}_1^*)\right)$$

$$\leq \sum_{i\in V_D}\log(1+n\,\mathrm{SNR}_s\sum_{k\in S}|\hat{H}_{ik}|^2) \tag{3.11}$$

$$\leq (\hat{v}-1)\sqrt{n}\log n\,\log(1+n^{2+\alpha(1/2+\delta)}\,\mathrm{SNR}_s) \tag{3.12}$$

w.h.p. for any $\delta>0$. Inequality (3.11) comes from the fact that for any covariance matrix Q of the transmissions from S, the SNR received by each node $i\in V_D$ is smaller than $n\,\mathrm{SNR}_s\hat{d}_i$. Inequality (3.12) is obtained by using the crude bound $\hat{d}_i\leq n^{1+\alpha(1/2+\delta)}$, which follows from the fact that the rescaled minimal separation between any two nodes in the network is larger than $\frac{1}{n^{1/2+\delta}}$ w.h.p. for any $\delta>0$ (Lemma 2.3.1-(a)) and the number of nodes in S are smaller than n. On the other hand, the number of nodes in V_D is upper bounded by $(\hat{v}-1)\sqrt{n}\log n$ w.h.p (Lemma 2.3.1-(b)).

The second term in (3.10) is the capacity of the MIMO channel between nodes in S and nodes in $D\backslash V_D$. The following lemma provides an upper bound on the capacity of this channel. Though the main idea is to upper bound the capacity by the total received SNR using inequality (3.6), this is not done immediately as we first need to waive out the possibility of communicating only through non-typically good channel matrices. Once (3.6) is applied, we need to handle the maximization over all admissible covariance matrices that are allowed to be functions of the channel state realizations.

Note that the upper bound given in the below lemma holds in general for any choice of $\hat{v}$, or equivalently $D\setminus V_D$. However, recall our earlier discussion that the upper bound will be tight only if the set $D\backslash V_D$ is tuned appropriately.

Lemma 3.3.1. *Let SNR_{tot} be the total SNR received by all the nodes in $D\backslash V_D$, when nodes in S are transmitting* independent *signals at full power, i.e.,*

$$SNR_{tot} := \sum_{i\in D\backslash V_D} SNR_i = SNR_s\sum_{i\in D\backslash V_D}\hat{d}_i. \tag{3.13}$$

Recall that SNR_i has been defined in (3.7) as the SNR received by the node $i\in D$ under independent transmissions from the left hand side. Then for every $\epsilon>0$,

$$\max_{\substack{Q(\hat{H}_2)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_2))\leq 1,\,\forall k\in S}} \mathbb{E}\left(\log\det(I+SNR_s\hat{H}_2Q(\hat{H}_2)\hat{H}_2^*)\right) \leq n^{\epsilon}SNR_{tot}. \tag{3.14}$$

Moreover, if $D \setminus V_D \neq \emptyset$ the scaling of the total received SNR can be evaluated to be

$$SNR_{tot} \leq \begin{cases} K_1\, SNR_s\, n^{2-\alpha/2} (\log n)^3 & 2 \leq \alpha < 3 \\ K_1\, SNR_s\, \hat{v}^{3-\alpha}\, \sqrt{n}\, (\log n)^3 & \alpha \geq 3. \end{cases} \tag{3.15}$$

w.h.p., where $K_1 > 0$ is a constant independent of SNR_s and n.

Lemma 3.3.1 says couple of surprising things. First of all, it says that independent signaling at the transmit nodes is sufficient to achieve the cutset upper bound. Note that, a priori, on the left-hand side of (3.14), nodes are allowed to cooperate and do any sort of transmit beamforming over channel state realizations. Lemma 3.3.1 says that this is not necessary. This explains why we earlier based our choice of $\hat{v}$ on the assumption of independent transmissions from the left-hand side nodes. Independent signalling is indeed good enough, at least as far as scaling of the capacity is concerned.

Second, depending on α, the lemma identifies a dichotomy on how the received SNR under independent transmissions scales with system size (3.15). This dichotomy can be interpreted as follows. The total received SNR is dominated either by the power transferred between node pairs separated by a relatively short-distance (of the order of $\hat{v}$) or by the power transferred between nodes far away (at distance of the order of $\sqrt{n}$). There are relatively fewer node *pairs* at distance $\hat{v}$ but the channels between these pairs are considerably stronger than the pairs at diameter distance. When the attenuation parameter α is less than 3, the received power is dominated by the transfer between the large number of node pairs at distance $\sqrt{n}$. There are n^2 node pairs separated by a rescaled distance of the order of $\sqrt{n}$, which yields a total SNR transfer of $\mathrm{SNR}_s \times n^2 \times \sqrt{n}^{-\alpha}$ between these pairs. This is the first term in (3.15) up to the logarithmic terms. When $\alpha \geq 3$, the received SNR in the cutset bound is dominated by the power transferred between node pairs at distance $\hat{v}$. There are an order of $\sqrt{n} \times \hat{v}^3$ pairs located at distance of the order of $\hat{v}$. (Consider the nodes in S located up-to $\hat{v}$ rescaled horizontal distance to the cut and those nodes in $D \setminus V_D$ located up-to $2\hat{v}$ horizontal distance to the cut. Then count the number of node pairs that are separated with a distance of the order of $\hat{v}$.) Hence, the total SNR transfer between these node pairs is equal to $\sqrt{n}\hat{v}^3 \times (\hat{v})^{-\alpha}$. This argument yields the second term in (3.15) up to the logarithmic terms.

Combining the upper bounds (3.12) and (3.14) together with our choices for $\hat{v}$ specified earlier, one can get an upper bound on $T_{L\to R}$ in terms of SNR_s and n. Here, we state the final result in terms of scaling exponents: Recall the definitions of the scaling exponents of the aggregate throughput in (2.8) and

the nearest neighbor SNR in (2.9). We have,

$$e(\alpha,\beta) \leq \begin{cases} 1 & \beta \geq \alpha/2-1 \\ 2-\alpha/2+\beta & \beta < \alpha/2-1 \text{ and } 2 \leq \alpha < 3 \\ 1/2+\beta & \beta \leq 0 \text{ and } \alpha \geq 3 \\ 1/2+\beta/(\alpha-2) & 0 < \beta < \alpha/2-1 \text{ and } \alpha \geq 3 \end{cases} \tag{3.16}$$

where we identify four different operating regimes depending on α and β.

Note that in the first regime the upper bound (3.12) is active with $\hat{v} = \sqrt{n}$ (or equivalently $V_D = D$) while (3.14) is zero. The capacity of the network is limited by the degrees of freedom in an $n \times n$ MIMO transmission between the left and the right hand side nodes. In the second regime, (3.14) with the corresponding upper bound being the first line in (3.15), yields a larger contribution than (3.12). The capacity is limited by the total received SNR in a MIMO transmission between the left-hand side nodes and $D \setminus V_D$. Note that this total received SNR is equal (in order) to the power transferred in a MIMO transmission between two groups of n nodes separated by a distance of the order of the diameter of the network, i.e., $n^2 \times (\sqrt{n})^{-\alpha} \times \mathrm{SNR}_s$.

In the third regime, (3.14) is active with $\hat{v} = 1$, or equivalently $V_D = \emptyset$ so (3.12) is zero. The corresponding upper bound is the second line in (3.15). The capacity in this regime is still limited by the total SNR received by nodes in $D \setminus V_D$ ($= D$ now) but the total is now dominated by the SNR transferred between the nearest nodes to the cut, i.e., $\sqrt{n}$ pairs separated by the nearest neighbor distance ($\hat{v} = 1$), yielding $\sqrt{n} \times \mathrm{SNR}_s$. Note that this is where we make use of the assumption that there are no nodes located at rescaled distance smaller than 1 to the cut. Due to this assumption, the choice $\hat{v} = 1$ vanishes the upper bound (3.12) and simultaneously yields $K_1 \mathrm{SNR}_s \sqrt{n} (\log n)^2$ in the last line in (3.15) for the total SNR transferred from S to D. If there were nodes closer than rescaled distance 1 to the cut, we would need to choose $\hat{v} < 1$ to vanish the contribution from (3.12) which would yield a larger value for the term $K_1 \mathrm{SNR}_s \hat{v}^{3-\alpha} \sqrt{n} (\log n)^2$. The difficulty is the following. We would like to conclude that in this regime, the power transfer between left and right-hand side nodes is dominated by the contribution of the order $\sqrt{n}$ nearest neighbor pairs located around the cut. However there can be a pair of nodes, one node on the left and the other one on the right of the cut, that is separated by a distance much smaller than the nearest neighbor distance in the network and the capacity of the channel between these two nodes can be much larger than the total contribution of the $\sqrt{n}$ nearest neighbor pairs. However, even though this may be the case for the cut considered, it is not possible to rely on such pairs for communicating inside the network, since these pairs do not form a path inside the network w.h.p. See how this fact is made precise in Appendix 3.B.

The most interesting regime is the fourth one. Both (3.12) and (3.14), with the choice $\hat{v} = \mathrm{SNR}_s^{\frac{1}{\alpha-2}}$, yield the same contribution. Note that (3.12) upper bounds the information transfer to V_D, the set of nodes that have bandwidth-limited connections to the left-hand side. This information transfer is limited

in degrees of freedom. On the other hand, (3.14) upper bounds the information transfer to $D \setminus V_D$, the set of nodes that have power-limited connections to the left-hand side. This second information transfer is power-limited. Therefore in this regime, the network capacity is both limited in degrees of freedom and power, since increasing the bandwidth increases the first term (3.12) and increasing the power increases the second term (3.14). This behavior is a consequence of the heterogeneous nature of links in a network and does not occur in point-to-point links. □

Proof of Lemma 3.3.1: We are interested in the scaling of the MIMO capacity,

$$\max_{\substack{Q\left(\hat{H}_2\right)\geq 0 \\ \mathbb{E}\left(Q_{kk}\left(\hat{H}_2\right)\right)\leq 1,\, \forall k\in S}} \mathbb{E}\left(\log\det\left(I+\mathrm{SNR}_s\hat{H}_2 Q\left(\hat{H}_2\right)\hat{H}_2^*\right)\right). \tag{3.17}$$

A natural way to upper bound (3.17) is to first relax the individual power constraint

$$\mathbb{E}\left(Q_{kk}\left(\hat{H}_2\right)\right)\leq 1,\, \forall k\in S$$

to a total power constraint,

$$\mathbb{E}\left(\mathrm{Tr}\, Q\left(\hat{H}_2\right)\right)\leq |S|$$

where $|S|$ denotes the cardinality of the set S. In the present context however, this is not convenient, since the matrix $\hat{H}_2$ is badly conditioned: some nodes in S are close to the cut and some are far apart, so the impact of their 1 Watt power on the system performance is quite different. A total transmit power constraint allows the transfer of power from the nodes far away from the cut to those nodes that are located close to the cut, resulting in a loose bound. Instead, we will relax the individual power constraints to a total *weighted* power constraint, where the weight assigned to a node is proportional to the impact of its unit power. The impact is measured by the total *received* power on the right-hand side of the cut per watt of transmit power from that left-hand side node.

Let us normalize the columns of the matrix $\hat{H}_2$ by dividing each column k by its norm. Let w_k denote the squared L^2-norm of the k'th column

$$w_k = \sum_{i\in D\setminus V_D} |\hat{H}_{ik}|^2,$$

We define the normalized matrix

$$\tilde{H}_{ik} = \frac{1}{\sqrt{w_k}}\,\hat{H}_{ik} \qquad i\in D\setminus V_D,\, k\in S. \tag{3.18}$$

The expression (3.17) is then equal to

$$\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\tilde{Q}_{kk}(\tilde{H}))\leq w_k,\,\forall k\in S}} \mathbb{E}\left(\log\det\left(I+\mathrm{SNR}_s\tilde{H}\tilde{Q}\left(\tilde{H}\right)\tilde{H}^*\right)\right).$$

Note that $\mathrm{SNR}_s w_k$ corresponds to the total received SNR by the nodes $D\setminus V_D$ of the signal sent by the user $k\in S$. Having weighted each of the individual power constraint in (3.17) by their impact, we now relax them to a total power constraint which yields the following upper bound for (3.17),

$$\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(H))\leq W_{tot}}} \mathbb{E}\left(\log\det(I+\mathrm{SNR}_s\tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\right),$$

where

$$W_{tot}=\sum_{k\in S}w_k=\sum_{k\in S,\,i\in D\setminus V_D}|\hat{H}_{ik}|^2.$$

Let us now define, for given $n\geq 1$ and $\varepsilon>0$, the event

$$B_{n,\varepsilon}=\{\|\tilde{H}\|^2>n^{\varepsilon}\},$$

where $\|A\|$ denotes the largest singular value of the matrix A. Note that the matrix $\tilde{H}$ is better conditioned than the original channel matrix $\hat{H}_2$: all the diagonal elements of $\tilde{H}\tilde{H}^*$ are roughly of the same order (up to a factor $\log n$), and it can be shown that there exists $K_2>0$ such that

$$\mathbb{E}(\|\tilde{H}\|^2)\leq K_2\,(\log n)^3$$

for all n. In Appendix 3.A, we show the following more precise statement.

Lemma 3.3.2. *For any $\varepsilon>0$ and $p\geq 1$, there exists $K_2>0$ such that for all n,*

$$\mathbb{P}(B_{n,\varepsilon})\leq\frac{K_2}{n^p}.$$

It follows that

$$\begin{aligned}
&\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\left(\log\det(I+\mathrm{SNR}_s\tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\right)\\
&\qquad\leq \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\left(\log\det(I+\mathrm{SNR}_s\tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\,1_{B_{n,\varepsilon}}\right)\\
&\qquad+ \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\left(\log\det(I+\mathrm{SNR}_s\tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\,1_{B^c_{n,\varepsilon}}\right) \quad (3.19)
\end{aligned}$$

The first term in (3.19) refers to the event that the channel matrix $\tilde{H}$ is accidentally ill-conditioned. Since the probability of such an event is polynomially small by Lemma 3.3.2, the contribution of this first term is actually negligible. In the second term in (3.19), the matrix $\tilde{H}$ is well conditioned, and this term is actually proportional to the maximum SNR transfer from S to $D \setminus V_D$. Details follow below.

For the first term in (3.19), we use Hadamard's inequality and obtain

$$\begin{aligned}
&\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\Big(\log\det(I + \mathrm{SNR}_s\, \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\, 1_{B_{n,\varepsilon}}\Big) \\
&\qquad = \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\Big(\log\det(I + \mathrm{SNR}_s\, \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*) \,\Big|\, B_{n,\varepsilon}\Big)\mathbb{P}(B_{n,\varepsilon}) \\
&\qquad \leq \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \mathbb{E}\Big(\sum_{i\in D\setminus V_D} \log(1 + \mathrm{SNR}_s\, \tilde{H}_i\tilde{Q}(\tilde{H})\tilde{H}_i^*) \,\Big|\, B_{n,\varepsilon}\Big)\mathbb{P}(B_{n,\varepsilon})
\end{aligned}$$

where $\tilde{H}_i$ is the i^{th} row of $\tilde{H}$. By the fact that

$$\tilde{H}_i\tilde{Q}(\tilde{H})\tilde{H}_i^* \;=\; \mathrm{Tr}\left(\tilde{H}_i\tilde{Q}(\tilde{H})\tilde{H}_i^*\right) \leq \|\tilde{H}_i\|^2\,\mathrm{Tr}\left(\tilde{Q}(\tilde{H})\right)$$

where $\|\tilde{H}_i\|^2$ is the squared norm of $\tilde{H}_i$, and using Jensen's inequality, this expression in turn is bounded above by

$$\begin{aligned}
&\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \sum_{i\in D\setminus V_D} \log\left(1 + \mathrm{SNR}_s\,\mathbb{E}\left(\|\tilde{H}_i\|^2\,\mathrm{Tr}\tilde{Q}(\tilde{H})\Big|\, B_{n,\varepsilon}\right)\right)\mathbb{P}(B_{n,\varepsilon}) \\
&\leq \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{tot}}} \sum_{i\in D\setminus V_D} \log\left(1 + \mathrm{SNR}_s\,\mathbb{E}\left(\|\tilde{H}_i\|^2\,\mathrm{Tr}\tilde{Q}(\tilde{H})\right)\Big/\mathbb{P}(B_{n,\varepsilon})\right)\mathbb{P}(B_{n,\varepsilon}) \\
&\leq n\log\left(1 + \mathrm{SNR}_s\,\frac{n\,W_{tot}}{\mathbb{P}(B_{n,\varepsilon})}\right)\,\mathbb{P}(B_{n,\varepsilon}).
\end{aligned}$$

The last inequality follows from upper-bounding $\|\tilde{H}_i\|^2$ as

$$\|\tilde{H}_i\|^2 = \sum_{k\in S} |\hat{H}_{ik}|^2\,\frac{1}{w_k} \leq \sum_{k\in S} 1 \leq n.$$

which follows from the definition of $\tilde{H}$ in (3.18). The fact that the rescaled minimum distance between the nodes in S and $D \setminus V_D$ is at least 1 yields

$$W_{tot} = \sum_{k\in S,\, i\in D\setminus V_D} |\hat{H}_{ik}|^2 < n^2.$$

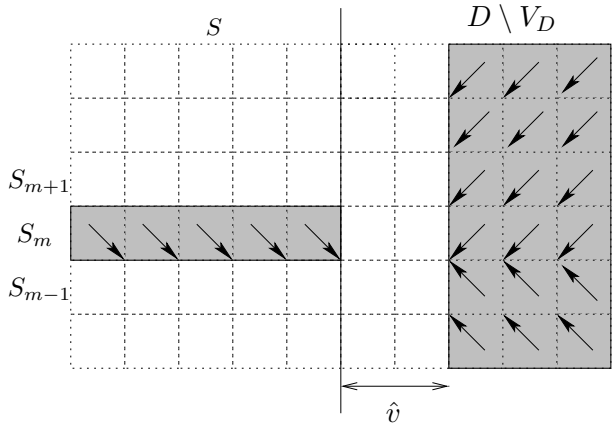

Figure 3.2: The displacement of the nodes inside the squarelets to squarelet vertices, indicated by arrows.

Noting that $x \mapsto x \log(1 + 1/x)$ is increasing on $[0, 1]$ and using Lemma 3.3.2, we obtain finally that for any $p \geq 1$, there exists $K_2 > 0$ such that

$$\max_{\substack{\tilde{Q}(\tilde{H}) \geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H})) \leq W_{tot}}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\, 1_{B_{n,\varepsilon}}\right) \leq K_2\, n^{1-p} \log\left(1 + \mathrm{SNR}_s\, \frac{n^{3+p}}{K_2}\right),$$

which decays polynomially to zero with arbitrary exponent as n tends to infinity.

For the second term in (3.19), we simply have

$$\begin{aligned}
\max_{\substack{\tilde{Q}(\tilde{H}) \geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H})) \leq W_{tot}}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\, 1_{B^c_{n,\varepsilon}}\right) & \\
\leq \max_{\substack{\tilde{Q}(\tilde{H}) \geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H})) \leq W_{tot}}} \mathbb{E}\left(\mathrm{Tr}(\mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)\, 1_{B^c_{n,\varepsilon}}\right) & \\
\leq \max_{\substack{\tilde{Q}(\tilde{H}) \geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H})) \leq W_{tot}}} \mathbb{E}\left(\mathrm{SNR}_s \|\tilde{H}\|^2 \mathrm{Tr}\tilde{Q}(\tilde{H})\, 1_{B^c_{n,\varepsilon}}\right) & \\
\leq n^{\varepsilon}\, \mathrm{SNR}_s W_{tot}. &
\end{aligned}$$

The last thing that needs therefore to be checked is the scaling of W_{tot} stated in Lemma 3.3.1.

Let us divide the rescaled network area of size n into n squarelets of area 1. By Part (b) of Lemma 2.3.1, there are no more than $\log n$ nodes in each

squarelet with high probability. Let us consider grouping the squarelets on the left of the cut into $\sqrt{n}$ rectangular areas of height 1 and width $\sqrt{n}/2$, as shown in Figure 3.2. Let S_m denote the nodes in S that are located on the m'th rectangle so that $S = \bigcup_{m=1}^{\sqrt{n}} S_m$. We are interested in bounding above

$$W_{tot} = \sum_{k \in S} w_k = \sum_{m=1}^{\sqrt{n}} \sum_{k \in S_m} w_k.$$

Let us consider

$$\sum_{k \in S_m} w_k = \sum_{k \in S_m, i \in D \setminus V_D} |\hat{H}_{ik}|^2 = \sum_{k \in S_m, i \in D \setminus V_D} \hat{r}_{ik}^{-\alpha} \tag{3.20}$$

for a given m. Note that if we move the points that lie in each squarelet of S_m together with the nodes in the squarelets of $D \setminus V_D$ onto the squarelet vertex as indicated by the arrows in Figure 3.2, all the (positive) terms in the summation in (3.20) can only increase since the displacement can only decrease the Euclidean distance between the nodes involved. Note that the modification results in a regular network with at most $\log n$ nodes at each squarelet vertex on the left and at most $2 \log n$ nodes at each squarelet vertex on the right. Considering the same reasoning for all rectangular slabs S_m, , $m = 1, \ldots, \sqrt{n}$ allows to conclude that W_{tot} for the random network is with high probability less than the same quantity computed for a regular network where nodes are located on a square grid of distance 1, with $\log n$ nodes at each left-hand side vertex and $2 \log n$ nodes at each right-hand side vertex.

The most convenient way to index the node positions in a regular network is to use double indices. The left-hand side nodes are located at positions $(-k_x, k_y)$ for $k_x = 0, \ldots, \sqrt{n}/2$, $k_y = 0, \ldots, \sqrt{n}$ and those on the right at positions (i_x, i_y) where $i_x = \hat{v}, \ldots, \sqrt{n}/2$ for $\hat{v} \geq 1$ and $i_y = 0, \ldots, \sqrt{n}$, so that

$$\hat{H}_{ik} = \frac{e^{j\,\theta_{ik}}}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/4}}$$

and

$$w_{k_x,k_y} = \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} \frac{1}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2}} \tag{3.21}$$

which yields the following upper bound for W_{tot} of the random network,

$$W_{tot} \leq 2(\log n)^2 \sum_{k_x=0}^{\sqrt{n}/2} \sum_{k_y=0}^{\sqrt{n}} w_{k_x,k_y}. \tag{3.22}$$

The following lemma establishes the scaling of w_{k_x,k_y} defined in (3.21).

Lemma 3.3.3. *There exist constants $K_3, K_4 > 0$ independent of k_x, k_y and n such that*

$$w_{k_x,k_y} \leq \begin{cases} K_3 \log n & \text{if } \alpha = 2, \\ K_3 \, (\hat{v} + k_x)^{2-\alpha} & \text{if } \alpha > 2, \end{cases}$$

and

$$w_{k_x,k_y} \geq K_4 \, (\hat{v} + k_x)^{2-\alpha} \qquad \text{for} \qquad \alpha \geq 2.$$

The rigorous proof of the lemma is given at the end of Appendix 3.A. A heuristic way of thinking about the approximation

$$w_{k_x,k_y} \approx (\hat{v} + k_x)^{2-\alpha}$$

can be obtained through Laplace's principle. The summation in w_{k_x,k_y} scales the same as the maximum term in the sum times the number of terms which have roughly this maximum value. The maximum term is of the order of $1/(\hat{v} + k_x)^{\alpha}$. The terms that take on roughly this value are those for which i_x runs from $\hat{v}$ to the order of $2\hat{v} + k_x$ and i_y runs from k_y to k_y plus or minus the order of $\hat{v} + k_x$. There are roughly $(\hat{v} + k_x)^2$ such terms. Hence $w_{k_x,k_y} \approx 1/(\hat{v} + k_x)^{\alpha} \cdot (\hat{v} + k_x)^2 = (\hat{v} + k_x)^{2-\alpha}$.

We can now use the upper bound given in Lemma 3.3.3 which gives

$$\sum_{k_x,k_y=0}^{\sqrt{n}} w_{k_x,k_y} \leq \begin{cases} K_5 \, n \log n & \text{if } \alpha = 2, \\ K_5 \, n^{2-\alpha/2} & \text{if } 2 < \alpha \leq 3, \\ K_5 \, \sqrt{n} \log n & \text{if } \alpha = 3, \\ K_5 \, \hat{v}^{3-\alpha} \sqrt{n} & \text{if } \alpha > 3 \end{cases}$$

for another constant $K_5 > 0$ independent of n. This upper bound combined with (3.22) yields (3.15) and completes the proof of Lemma 3.3.1. □

3.4 Discussion

In the next chapter, we search for communication schemes whose performance meets the upper bound derived in the current chapter. In this section, we would like to summarize the insights provided by the upper bound derivation on the properties of such optimal schemes.

In the first two regimes, we have seen that the capacity of the network is limited by the degrees of freedom and received SNR respectively, in a network wide MIMO transmission. Hence, we expect the optimal schemes for these regimes to imitate such cooperative MIMO transmissions. However at this point, it is not at all obvious that such cooperative MIMO transmissions can be realized efficiently in a distributed network setup. Note that in the derivation of the upper bound, we have assumed that the transmitting nodes in S and the receiving nodes in D can cooperate for free among themselves. In reality, establishing cooperation between these nodes may be overwhelming. We will see in Section 4.4 that this is not the case, as efficient architectures for establishing cooperation can be devised.

In the third regime, the information transfer between the two halves of the network is limited by the power transferred between the closest nodes to the cut. This observation suggests the following idea: if the objective is to transfer information from the left-half network to the right-half, it is enough to employ only those pairs that are located closest to the cut and separated by the nearest neighbor distance. The rest of the nodes in the network can undertake simultaneous transmissions, suggesting the idea of spatial reuse. In other words, the upper bound derivation suggests that efficient transmissions in this regime are the point-to-point transmissions between nearest neighbors. Indeed, this is how the well-known nearest neighbor multi-hopping scheme transfers power across the network, so the multi-hopping scheme arises as a natural candidate for optimality in the third regime.

In the derivation of the upper bound for the fourth regime, we have seen that the two terms (3.12) and (3.14), governing the information transfer from S to V_D and $D \setminus V_D$ respectively, yield the same contribution with the particular choice $\hat{w} = \mathrm{SNR}_s^{\frac{1}{\alpha-2}}$. Since the contributions of the two terms are equal (and since we are interested in scaling here) the derivation of the upper bound suggests the following idea: information can be transferred optimally from the left-half network to the right-half by performing MIMO transmission only between those nodes on both sides of the cut that are located up to $\hat{w} = \mathrm{SNR}_s^{\frac{1}{\alpha-2}}$ rescaled distance to the cut. Note that (3.12) corresponds to the degrees of freedom in such a MIMO transmission. As in the case of multi-hop, we can have spatial reuse and allow the rest of the nodes in the network to perform simultaneous transmissions. Thus, the derivation of the upper bound suggests that efficient transmissions in the fourth regime are MIMO transmissions at the scale $\hat{w} = \mathrm{SNR}_s^{\frac{1}{\alpha-2}}$. Combined with the idea of spatial reuse, this understanding suggests to transfer information in the network by performing MIMO transmissions at the particular (local) scale of $\hat{w} = \mathrm{SNR}_s^{\frac{1}{\alpha-2}}$ and then multi-hopping at the global scale. The MIMO transmissions are again based on the cooperation architecture that will be presented Section 4.4. This hybrid scheme combining MIMO with multi-hopping will be introduced in Section 4.6.

3.A Largest Eigenvalue of the Equalized Channel Matrix $\tilde{H}$

In this appendix, we give the proofs of Lemma 3.3.2 and Lemma 3.3.3. We start with Lemma 3.3.2. The proof of the second lemma is given at the end of the section.

Proof of Lemma 3.3.2: Let us start by considering the $2m^{th}$ moment of the spectral norm of $\tilde{H}$ given by (see [29, Ch. 5])

$$\|\tilde{H}\|^{2m} = \rho(\tilde{H}^*\tilde{H})^m = \lim_{l\to\infty} \{\mathrm{Tr}((\tilde{H}^*\tilde{H})^l)\}^{m/l},$$

where $\rho(A)$ is the largest eigenvalue of the positive-semi-definite matrix A which is called the spectral radius of A. By dominated convergence theorem and Jensen's inequality, we have

$$\mathbb{E}(\|\tilde{H}\|^{2m}) \leq \lim_{l\to\infty} \{\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^l))\}^{m/l}.$$

In the subsequent paragraphs, we will prove that the following upper bound holds with high probability,

$$\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^l)) \leq t_l\, n\, (K_2' \log n)^{3l} \tag{3.23}$$

where $t_l = \frac{(2l)!}{l!(l+1)!}$ are the Catalan numbers and $K_2' > 0$ is a constant independent of n. By Chebyshev's inequality, this allows to conclude that for any m,

$$\begin{aligned}
\mathbb{P}(B_{n,\varepsilon}) &= \mathbb{P}\left(\|\tilde{H}\|^2 > n^{\epsilon}\right) \\
&\leq \frac{\mathbb{E}(\|\tilde{H}\|^{2m})}{n^{m\varepsilon}} \\
&\leq \frac{1}{n^{m\varepsilon}} \lim_{l\to\infty} (t_l\, n\, (K_1' \log n)^{3l})^{m/l} \\
&\leq \frac{(4(K_2' \log n)^3)^m}{n^{m\varepsilon}},
\end{aligned}$$

since $\lim_{l\to\infty} t_l^{1/l} = 4$. For any $\varepsilon > 0$, choosing m sufficiently large shows therefore that $\mathbb{P}(B_{n,\varepsilon})$ decays polynomially with arbitrary exponent as $n \to \infty$, which is the result stated in Lemma 3.3.2.

There remains to prove the upperbound in (3.23). Expanding the expression gives

$$\begin{aligned}
&\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^l)) \\
&= \sum_{\substack{i_1,\ldots,i_l \in D\setminus V_D \\ k_1\ldots,k_l \in S}} \mathbb{E}\left(\overline{\tilde{H}_{i_1k_1}}\tilde{H}_{i_1k_2}\overline{\tilde{H}_{i_2k_2}}\tilde{H}_{i_2k_3}\ldots\overline{\tilde{H}_{i_lk_l}}\tilde{H}_{i_lk_1}\right).
\end{aligned} \tag{3.24}$$

Recall that the random variables $\tilde{H}_{ik}$ are independent and zero-mean, so the expectation is only non-zero when the terms in the product form conjugate pairs. Therefore, in order to upper bound (3.24) we first need to identify the nonzero terms in the summation.

Let us consider the case $l = 2$ as an example. We have,

$$\begin{aligned}
&\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^2)) \\
&= \sum_{\substack{i_1,i_2 \in D\setminus V_D \\ k_1,k_2 \in S}} \mathbb{E}\left(\overline{\tilde{H}_{i_1k_1}}\tilde{H}_{i_1k_2}\overline{\tilde{H}_{i_2k_2}}\tilde{H}_{i_2k_1}\right) &(3.25)\\
&= \sum_{\substack{i_1,i_2 \in D\setminus V_D \\ k \in S}} |\tilde{H}_{i_1k}|^2|\tilde{H}_{i_2k}|^2 + \sum_{\substack{i \in D\setminus V_D \\ k_1\neq k_2 \in S}} |\tilde{H}_{ik_1}|^2|\tilde{H}_{ik_2}|^2 &(3.26)
\end{aligned}$$

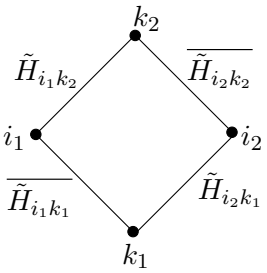

Figure 3.3: The product in Eq. 3.25 illustrated as a ring.

since the expectation is non-zero only when either $k_1 = k_2 = k$ or $i_1 = i_2 = i$. Note that we have removed the expectations in (3.26) since $|\tilde{H}_{ik}|^2$ is a deterministic quantity in our case. The expression can be bounded above by

$$\begin{aligned} &\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^2)) \\ &\leq \sum_{\substack{i_1,i_2\in D\setminus V_D \\ k\in S}} |\tilde{H}_{i_1k}|^2|\tilde{H}_{i_2k}|^2 + \sum_{\substack{i\in D\setminus V_D \\ k_1,k_2\in S}} |\tilde{H}_{ik_1}|^2|\tilde{H}_{ik_2}|^2 \end{aligned} \tag{3.27}$$

where we now double-count the terms with $i_1 = i_2 = i$ and $k_1 = k_2 = k$, that is, the terms of the form $|\tilde{H}_{ik}|^4$.

The non-vanishing terms in the sum in (3.25) can also be determined by the following approach, which generalizes to larger l: let each index be associated to a vertex and each term in the product in (3.25) to an edge between its corresponding vertices. Note that the resulting graph is in general a ring with 4 edges as depicted in Figure 3.3. A term in the summation in (3.25) is only non-zero if each edge of its corresponding graph has even multiplicity. Such a graph can be obtained from the ring in Figure 3.3 by merging some of the vertices, thus equating their corresponding indices. For example, merging the vertices k_1 and k_2 into a single vertex k gives the graph in Figure 3.4-a; on the other hand, merging i_1 and i_2 into a single vertex i gives Figure 3.4-b. Note that in the first figure i_1, i_2 can take values in $D \setminus V_D$ and k can take values in S, thus the sum of all such terms yields

$$\sum_{\substack{i_1,i_2\in D\setminus V_D \\ k\in S}} |\tilde{H}_{i_1k}|^2|\tilde{H}_{i_2k}|^2. \tag{3.28}$$

Similarly, the terms of the form in Figure 3.4-b sum up to

$$\sum_{\substack{i\in D\setminus V_D \\ k_1,k_2\in S}} |\tilde{H}_{ik_1}|^2|\tilde{H}_{ik_2}|^2. \tag{3.29}$$

Note that another possible graph composed of edges with even multiplicity can be obtained by further merging the vertices i_1 and i_2 into a single vertex i

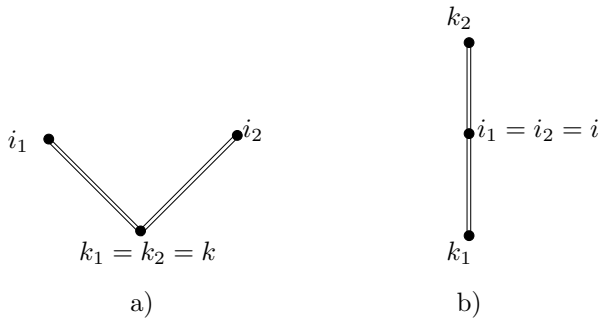

Figure 3.4: Two possible graphs corresponding to the non-zero terms in (3.25).

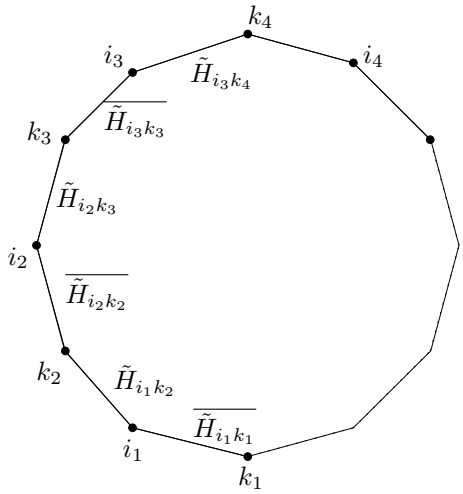

Figure 3.5: The product in Eq. 3.24 illustrated as a ring.

in Figure 3.4-a, or equivalently merging k_1 and k_2 into k in Figure 3.4-b. This will result in a graph with only two vertices k and i and a quadruple edge in between which corresponds to terms of the form $|\tilde{H}_{ik}|^4$ with $i \in D \setminus V_D$ and $k \in S$. Note however that such terms have already been considered in both (3.28) and (3.29) since we did not exclude the case $i_1 = i_2$ in (3.28) and $k_1 = k_2$ in (3.29). In fact, terms corresponding to any graph with number of vertices less than 3 are already accounted for in either one of the sums in (3.28) and (3.29), or simultaneously in both. Hence, the sum of (3.28) and (3.29) is an upper bound for (3.25) yielding again (3.27).

In the general case with $l \geq 2$, considering (3.24) leads to a larger ring with $2l$ edges, as depicted in Figure 3.5. Similarly to the case $l = 2$, the non-vanishing terms in (3.24) are those that correspond to a graph having only edges of even multiplicity. Since each edge can have at least double multiplicity, such graphs can have at most l edges. In turn, a graph with l edges can have at most $l + 1$ vertices which is the case of a tree. Hence, let us

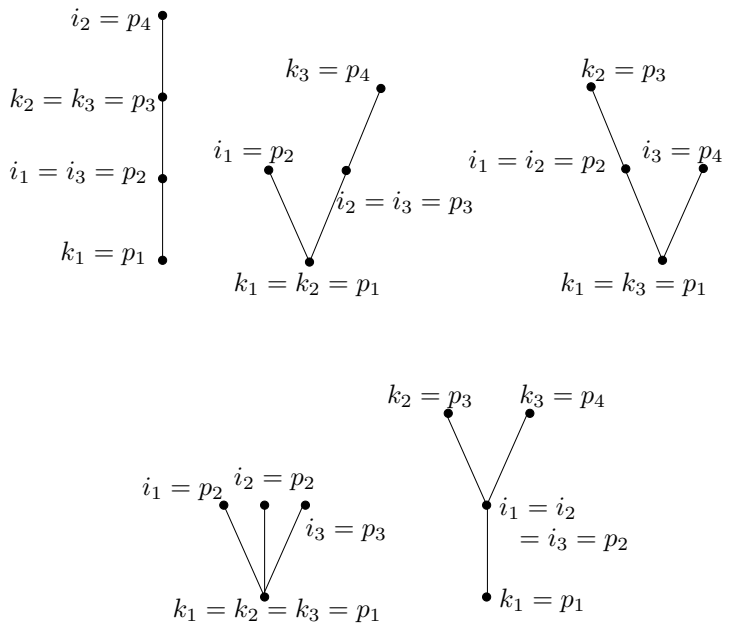

Figure 3.6: Planar rooted planted trees with 3 branches. Note that each edge is actually a double edge in our case, although depicted with a single line in the figure.

first start by considering such trees; namely, planar trees with l branches that are rooted (at k_1) and planted, implying that rotating asymmetric trees around the root results in a new tree. See Figure 3.6 which depicts the five possible trees with $l = 3$ branches where we relabel the resultant $l + 1 = 4$ vertices as $p_1, \ldots, p_4$. In general, the number of different planar, rooted, planted trees with l branches is given by the l'th Catalan number t_l [48]. In each of these trees, the $l + 1$ vertices $p_1, \ldots, p_{l+1}$ take values in either $D \setminus V_D$ or S. Hence, each tree $\mathscr{T}_q^l,\ q = 1, \ldots, t_l$ corresponds to a group of non-zero terms,

$$T_q^l = \sum_{p_1,\ldots,p_{l+1}} f_{\mathscr{T}_q^l}(p_1, \ldots, p_{l+1}), \qquad q = 1, \ldots, t_l. \tag{3.30}$$

Note that if a non-vanishing term in (3.24) corresponds to a graph with less than $l + 1$ vertices, then the corresponding graph possesses either edges with multiplicity larger than 2 or cycles, and this term is already accounted for in either one or more of the terms in (3.30). This fact can be observed by noticing that both edges with large multiplicity as well as cycles can be untied to get trees with l branches, with some of the $l + 1$ indices constrained however to share the same values (see Figure 3.7). Note that such cases are not excluded from the summations in (3.30), thus we have

$$\mathbb{E}(\mathrm{Tr}((\tilde{H}^*\tilde{H})^l)) \leq \sum_{q=1}^{t_l} T_q^l. \tag{3.31}$$

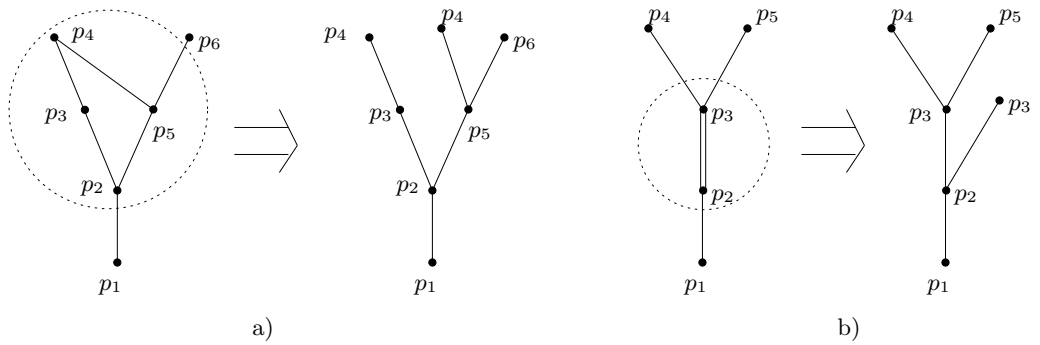

Figure 3.7: The graphs with a) cycles and b) edges with large multiplicity, can be untied to get trees with some vertices constrained to share the same indices.

Now that we have identified the non-zero terms in the summation in (3.24), we need to evaluate the order of these terms. Below we show that

$$T_q^l \leq n(K_2' \log n)^l, \qquad \forall q \tag{3.32}$$

in a regular network and

$$T_q^l \leq n(K_2' \log n)^{3l}, \qquad \forall q \tag{3.33}$$

with high probability in a random network for a constant $K_2' > 0$ independent of n. We first concentrate on regular networks in order to reveal the proof idea in the simplest setting. A binning argument then allows to extend the result to random networks.

a) Regular network: In the regular case, we assume the nodes on the left-half domain S are located at positions $(-k_x, k_y)$ for $k_x = 0, \ldots, \sqrt{n}/2$, $k_y = 0, \ldots, \sqrt{n}$ and those on the right-half domain D at (i_x, i_y) for $i_x = 1, \ldots, \sqrt{n}/2$, $i_y = 0, \ldots, \sqrt{n}$. Recall that the matrix $\tilde{H}$ concerns the nodes in S and nodes in $D \setminus V_D$. Therefore, the elements of the matrix $\tilde{H}$ corresponding to a regular network are given by

$$\tilde{H}_{ik} = \frac{e^{j\,\theta_{ik}}}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/4}} \frac{1}{\sqrt{w_{k_x,k_y}}},$$

for $k_x = 0, \ldots, \sqrt{n}/2$, $k_y = 0, \ldots, \sqrt{n}$ and $i_x = \hat{v}, \ldots, \sqrt{n}/2$, $i_y = 0, \ldots, \sqrt{n}$, where

$$w_{k_x,k_y} = \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} \frac{1}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2}}.$$

In the discussion below, we will need an upper bound on the scaling of

$$\sum_{i \in D \setminus V_D} |\tilde{H}_{ik}|^2 = \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} |\tilde{H}_{ik}|^2 \text{ and } \sum_{k \in S} |\tilde{H}_{ik}|^2 = \sum_{k_x=0}^{\sqrt{n}/2} \sum_{k_y=0}^{\sqrt{n}} |\tilde{H}_{ik}|^2.$$

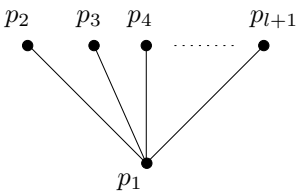

Figure 3.8: A simple tree with l branches.

By Lemma 3.3.3, we have

$$w_{k_x,k_y} \geq K_4 \, (\hat{v} + k_x)^{2-\alpha}$$

for a constant $K_4 > 0$ independent of n which, in turn, yields the upper bound

$$\begin{aligned} |\tilde{H}_{ik}|^2 &\leq \frac{1}{K_4} \frac{(\hat{v} + k_x)^{\alpha-2}}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2}} \\ &\leq \frac{1}{K_4} \frac{(\hat{v} + k_x)^{\alpha-2}}{(i_x + k_x)^2 + (i_y - k_y)^2} \frac{(\hat{v} + k_x)^{\alpha-2}}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2-1}} \\ &\leq \frac{1}{K_4} \frac{1}{(i_x + k_x)^2 + (i_y - k_y)^2}, \end{aligned} \tag{3.34}$$

where we make use of the fact that $i_x \geq \hat{v}$. Summing (3.34) over either i or k, and using the upper bound in Lemma 3.3.3 for $\alpha = 2$ yields

$$\sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} |\tilde{H}_{ik}|^2, \quad \sum_{k_x=0}^{\sqrt{n}/2} \sum_{k_y=0}^{\sqrt{n}} |\tilde{H}_{ik}|^2 \leq K_2' \log n \tag{3.35}$$

where $K_2' = \frac{K_3}{K_4}$ with K_3 and K_4 being the constants appearing in the lemma.

Let us now go back to evaluating the upper bound in (3.31). Let us first consider the simplest case where the tree is composed of l height 1 branches and denote it by $\mathscr{T}_1^l$ (see Figure 3.8). We have

$$\begin{aligned} T_1^l &= \sum_{p_1,\ldots,p_{l+1}} f_{\mathscr{T}_1^l}(p_1, \ldots, p_{l+1}) \\ &= \sum_{p_1,\ldots,p_{l+1}} |\tilde{H}_{p_2p_1}|^2 \, |\tilde{H}_{p_3p_1}|^2 \ldots |\tilde{H}_{p_{l+1}p_1}|^2 \\ &= \sum_{p_1 \in S} \left(\sum_{p_2 \in D \setminus V_D} |\tilde{H}_{p_2p_1}|^2 \right)^l \\ &\leq n(K_2' \log n)^l \end{aligned} \tag{3.36}$$

which follows from the upper bound (3.35).

Now let us consider the general case of an arbitrary tree $\mathscr{T}_q^l$ having s leaves, where $1 \leq s \leq l$ (see Figure 3.9). Let the indices corresponding to these leaves be $m_1, \ldots, m_s$. Let us denote the "parent" vertices of these leaves by $p_1, \ldots, p_{s'}$. Note that $s' \leq s$ since some leaves may be sharing the same "parent" vertex. Assume that p_1 is the common parent vertex of leaves $m_1, \ldots, m_{d_1}$; p_2 is the common parent vertex of leaves $m_{(d_1+1)}, \ldots, m_{d_2}$ etc. and finally $p_{s'}$ is the parent of $m_{(d_t+1)}, \ldots, m_s$. The term T_q^l corresponding to this tree is given by

$$\begin{aligned} T_q^l &= \sum_{\substack{m_1,\ldots,m_s \\ p_1,\ldots,p_{(l+1-s)}}} f_{\mathscr{T}_q^l}(p_1, \ldots, p_{l+1-s}, m_1, \ldots, m_s) \\ &= \sum_{p_1,\ldots,p_{(l+1-s)}} f_{\mathscr{T}_{q'}^{l-s}}(p_1, \ldots, p_{(l+1-s)}) \\ &\quad \times \sum_{m_1,\ldots,m_s} |\tilde{H}_{m_1 p_1}|^2 \ldots |\tilde{H}_{m_{d_1} p_1}|^2 \ |\tilde{H}_{m_{(d_1+1)} p_2}|^2 \ldots |\tilde{H}_{m_{d_2} p_2}|^2 \\ &\qquad \times \ldots |\tilde{H}_{m_{(d_t+1)} p_{s'}}|^2 \ldots |\tilde{H}_{m_s p_{s'}}|^2 \qquad (3.37) \\ &\leq T_{q'}^{l-s} (K_2' \log n)^s \qquad (3.38) \end{aligned}$$

where $T_{q'}^{l-s}$ corresponds to a smaller (and shorter) tree $\mathscr{T}_{q'}^{(l-s)}$ with $l - s$ branches.[3] The observation made in (3.38) is that the expression

$$f_{\mathscr{T}_q^l}(p_1, \ldots, p_{l+1-s}, m_1, \ldots, m_s)$$

depends on a leaf index m only through a single term of the form $|\tilde{H}_{mp}|^2$ hence the summation $\sum_m |\tilde{H}_{mp}|^2$ can be readily evaluated. Once the terms corresponding to all leaves of a parent node p are separated from the expression $f_{\mathscr{T}_q^l}(p_1, \ldots, p_{l+1-s}, m_1, \ldots, m_s)$, in the remaining expression

$$f_{\mathscr{T}_{q'}^{l-s}}(p_1, \ldots, p_{(l+1-s)}),$$

p is a leaf. The argument above decreases the height of the tree by 1, hence can be applied recursively to get a simple tree composed only of height 1 branches in which case the upper bound in (3.36) applies. Thus, given $\mathscr{T}_q^l$ let h be the number of recursions to get a simple tree and $s_1, \ldots, s_h$ denote the number of leaves in the trees observed at each step of the recursion. We have

$$\begin{aligned} T_q^l &\leq (K_2' \log n)^{s_1} (K_2' \log n)^{s_2} \ldots (K_2' \log n)^{s_h} T_1^{l-s_1 \cdots - s_h} \\ &\leq n (K_2' \log n)^l \end{aligned}$$

since $T_1^{l-s_1 \cdots - s_h} \leq n (K_2' \log n)^{l-s_1 \cdots - s_h}$ by (3.36). Thus, (3.32) follows.

[3]Note that the term corresponding to a leaf m can be either $|\tilde{H}_{mp}|^2$ or $|\tilde{H}_{pm}|^2$ depending on whether the height of the leaf is even or odd. However, in (3.37), we ignore this issue in order to simplify the notation since the upper bound (3.38) applies in both cases.

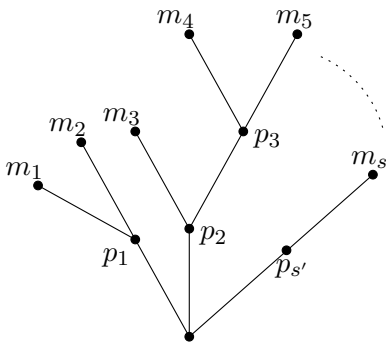

Figure 3.9: A tree with leaves $m_1, m_2, \ldots, m_s$.

b) Random network: We denote the locations of the nodes to the left of the cut by $a_k = (-a_k^x, a_k^y)$ where a_k^x is the x-coordinate and a_k^y is the y-coordinate of node $k \in S$ and those to the right of the cut are similarly denoted by $b_i = (b_i^x, b_i^y)$ for $i \in D \setminus V_D$. In this case, the matrix elements of $\tilde{H}$ are given by

$$\tilde{H}_{ik} = \frac{e^{j\,\theta_{ik}}}{((b_i^x + a_k^x)^2 + (b_i^y - a_k^y)^2)^{\alpha/4}} \frac{1}{\sqrt{w_k}}$$

and

$$w_k = \sum_{i \in D \setminus V_D} \frac{1}{((b_i^x + a_k^x)^2 + (b_i^y - a_k^y)^2)^{\alpha/2}}.$$

In parallel to the regular case, we will need an upper bound on $\sum_{i \in D \setminus V_D} |\tilde{H}_{ik}|^2$ and $\sum_{k \in S} |\tilde{H}_{ik}|^2$. The upper bound can be obtained in two steps by first showing that

$$w_k \geq K_4 \frac{(\hat{v} + a_k^x)^{2-\alpha}}{\log n} \tag{3.39}$$

with high probability for a constant $K_4 > 0$ independent of n, which leads to

$$\begin{aligned} |\tilde{H}_{ik}|^2 &\leq \frac{1}{K_4} \log n \, \frac{(\hat{v} + a_k^x)^{\alpha-2}}{((b_i^x + a_k^x)^2 + (b_i^y - a_k^x)^2)^{\alpha/2}} \\ &\leq \frac{1}{K_4} \log n \, \frac{1}{(b_i^x + a_k^x)^2 + (b_i^y - a_k^y)^2} \end{aligned} \tag{3.40}$$

for all $i \in S$, $k \in D \setminus V_D$. This, in turn yields

$$\sum_{k \in S} |\tilde{H}_{ik}|^2, \sum_{i \in D \setminus V_D} |\tilde{H}_{ik}|^2 \leq K_2' (\log n)^3 \tag{3.41}$$

with high probability for another constant $K_2' > 0$ independent of n. Recalling the leaf removal argument discussed for regular networks immediately leads to (3.33).

Both the lower bound in (3.39) and the upper bound in (3.41) regarding random networks can be proved using binning arguments that provide the connection to regular networks. In order to prove the lower bound, we consider Part (d) of Lemma 2.3.1, while the upper bound (3.41) is proved using Part (c) of the same lemma.

Let us first consider dividing the right-half network into squarelets of area $2\log n$. Given a left-hand side node k located at $(-a_k^x, a_k^y)$, let us move the nodes inside each right-hand side squarelet onto the squarelet vertex that is farthest to k. Since this displacement can only increase the Euclidean distance between the nodes involved, and since by Part (d) of Lemma 2.3.1, we know that there is at least one node inside each squarelet, we have

$$\begin{aligned} w_k &= \sum_{i\in D\setminus V_D} \frac{1}{((b_i^x + a_k^x)^2 + (b_i^y - a_k^y)^2)^{\alpha/2}} \\ &\geq \sum_{i_x=\hat{v}/\sqrt{2\log n}+1}^{\sqrt{n/8\log n}} \sum_{i_y=0}^{\sqrt{n/2\log n}} \frac{1}{((i_x\sqrt{2\log n} + a_k^x)^2 + (i_y\sqrt{2\log n} - a_k^x)^2)^{\alpha/2}} \\ &\geq K_4 \frac{(\hat{v} + a_k^x)^{2-\alpha}}{2\log n} \end{aligned}$$

by using the lower bound in Lemma 3.3.3.

Now having (3.40) in hand, in order to show (3.41), we divide the network into n squarelets of area 1. By Part (c) of Lemma 2.3.1, there are at most $\log n$ nodes inside each squarelet. Considering the argument in Section 3.3 and the displacement of the nodes as illustrated in Figure 3.2 yields a regular network with at most $2\log n$ nodes at each vertex in the right-half network,

$$\begin{aligned} \sum_{i\in D\setminus V_D} |\tilde{H}_{ik}|^2 &\leq \frac{2}{K_4}\log n \sum_{i\in D\setminus V_D} \frac{1}{(b_i^x + a_k^x)^2 + (b_i^y - a_k^y)^2} \\ &\leq \frac{4}{K_4}(\log n)^2 \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} \frac{1}{(i_x + k_x)^2 + (i_y - k_y)^2} \\ &\leq 4K_2'(\log n)^3. \end{aligned}$$

by employing the upper bound in Lemma 3.3.3 for $\alpha = 2$. The same bound follows similarly for $\sum_{k\in S} |\tilde{H}_{ik}|^2$, and hence the desired result in (3.41). □

Proof of Lemma 3.3.3: Both the lower and upper bound for w_{k_x,k_y} can be obtained by straightforward manipulations. Recall that

$$w_{k_x,k_y} = \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} \frac{1}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2}}.$$

The upper bound can be obtained as follows:

$$\begin{aligned}
w_{k_x,k_y} &= \sum_{y=-k_y}^{\sqrt{n}-k_y} \sum_{x=\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}} \\
&\leq \sum_{y=-k_y}^{\sqrt{n}-k_y} \left(\frac{1}{((\hat{v}+k_x)^2+y^2)^{\alpha/2}} + \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}}\, dx \right) \\
&\leq (\hat{v}+k_x)^{-\alpha} + \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{x^\alpha} dx + \int_{-k_y}^{\sqrt{n}-k_y} \frac{1}{((\hat{v}+k_x)^2+y^2)^{\alpha/2}}\, dy \\
&\quad + \int_{-k_y}^{\sqrt{n}-k_y} \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}}\, dx\, dy \\
&\leq (\hat{v}+k_x)^{-\alpha} + (1+\pi)(\hat{v}+k_x)^{1-\alpha} + \int_{-\pi/2}^{\pi/2} \int_{\hat{v}+k_x}^{\sqrt{2n}} \frac{1}{r^\alpha} r\, dr\, d\theta.
\end{aligned}$$

So

$$\begin{aligned}
w_{k_x,k_y} &= \begin{cases} (\hat{v}+k_x)^{-\alpha} + (1+\pi)(\hat{v}+k_x)^{1-\alpha} + \pi \log r \Big|_{\hat{v}+k_x}^{\sqrt{2n}} & \text{if } \alpha = 2, \\ (\hat{v}+k_x)^{-\alpha} + (1+\pi)(\hat{v}+k_x)^{1-\alpha} + \frac{\pi}{(2-\alpha)} r^{2-\alpha} \Big|_{\hat{v}+k_x}^{\sqrt{2n}} & \text{if } \alpha > 2, \end{cases} \qquad (3.42) \\
&\leq \begin{cases} K_3 \log n & \text{if } \alpha = 2, \\ K_3\, (\hat{v}+k_x)^{2-\alpha} & \text{if } \alpha > 2, \end{cases}
\end{aligned}$$

for a constant $K_3 > 0$ independent of n, since the dominating terms in (3.42) are the third ones.

The lower bound follows similarly:

$$\begin{aligned}
w_{k_x,k_y} &= \sum_{y=-k_y}^{\sqrt{n}-k_y} \sum_{x=\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}} \\
&\geq \sum_{y=-k_y}^{\sqrt{n}-k_y} \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}}\, dx \\
&\geq \int_{-k_y}^{\sqrt{n}-k_y} \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}}\, dx\, dy \quad - \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{x^\alpha}\, dx
\end{aligned}$$

thus,

$$\begin{aligned}
w_{k_x,k_y} &\geq \int_0^{\sqrt{n}} \int_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \frac{1}{(x^2+y^2)^{\alpha/2}}\, dx\, dy \quad + \frac{x^{1-\alpha}}{\alpha-1} \Big|_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} \\
&\geq \int_0^{\arctan(1/2)} \int_{\sqrt{2}(\hat{v}+k_x)}^{k_x+\sqrt{n}/2} \frac{1}{r^\alpha}\, r\, dr\, d\theta \quad + \frac{x^{1-\alpha}}{\alpha-1} \Big|_{\hat{v}+k_x}^{k_x+\sqrt{n}/2}.
\end{aligned}$$

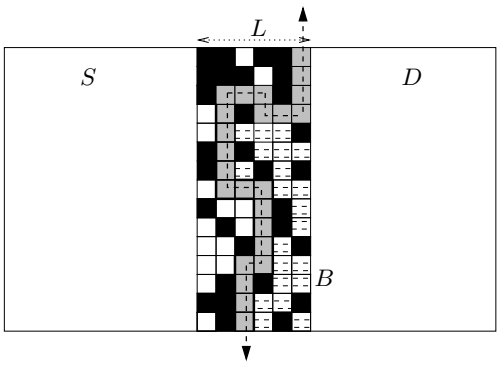

Figure 3.10: The cut in Lemma 3.B.1 that is free of nodes on both sides up to distance $c/2$ is illustrated in the figure.

So for all $\alpha \geq 2$, we have

$$w_{k_x,k_y} = \begin{cases} \arctan(\frac{1}{2}) \log r \Big|_{\sqrt{2}(\hat{v}+k_x)}^{k_x+\sqrt{n}/2} + \frac{1}{\alpha-1} x^{1-\alpha} \Big|_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} & \text{if } \alpha = 2, \\ \arctan(\frac{1}{2}) \frac{1}{2-\alpha} r^{2-\alpha} \Big|_{\sqrt{2}(\hat{v}+k_x)}^{k_x+\sqrt{n}/2} + \frac{1}{\alpha-1} x^{1-\alpha} \Big|_{\hat{v}+k_x}^{k_x+\sqrt{n}/2} & \text{if } \alpha > 2, \end{cases} \tag{3.43}$$
$$\geq K_4 (\hat{v} + k_x)^{2-\alpha}$$

where $K_4 > 0$ is a constant independent of n, since the dominating terms in (3.43) are the first ones. This concludes the proof of the lemma. □

3.B Removing Assumption 3.3.1

While proving the upper bound on network capacity in Section 3.3, we have considered a vertical cut of the network that divides the network area into two equal halves and assumed that there is an empty rectangular region to the right of this cut, of width equal to the nearest neighbor distance in the network (or of width equal to 1 in the corresponding rescaled network). With high probability, this assumption does not hold in a random realization of the network. Indeed for any linear cut of the random network, w.h.p. there will be nodes on both sides of the cut that are located at a distance much smaller than the nearest neighbor distance to the cut. In order to prove the result in Section 3.3 rigorously for random networks, we need to consider a cut that is not necessarily linear but satisfies the property of having no nodes located closer than the nearest neighbor distance to it. Below, we show the existence of such a cut using methods from percolation theory. See [18] for a more general discussion of applications of percolation theory to wireless networks.

Lemma 3.B.1. *For any realization of the random network and a constant $0 < c < 1/7\sqrt{e}$ independent of n and A, w.h.p. there exists a vertical cut of*

the network area that is not necessarily linear but is located in the middle of the network in a slab not wider than $L = c\sqrt{A/n}\log n$ *and is such that there exists no nodes at distance smaller than* $\frac{c}{2}\sqrt{A/n}$ *to the cut on both sides. See Fig. 3.10.*

The assumption of an empty region E in Section 3.3, allowed us to plug in $\hat{v} = 1$ in the second line of (3.15) and conclude that when the left-hand side nodes S are transmitting independent signals, the total SNR received by *all nodes* D to the right of the linear cut is bounded above by

$$\begin{aligned} \mathrm{SNR}_{tot} &= \sum_{i\in D} \mathrm{SNR}_i \\ &\leq \begin{cases} K_1\,\mathrm{SNR}_s\, n^{2-\alpha/2}(\log n)^3 & 2 \leq \alpha < 3 \\ K_1\,\mathrm{SNR}_s\,\sqrt{n}\,(\log n)^3 & \alpha \geq 3, \end{cases} \end{aligned} \tag{3.44}$$

where SNR_i is defined in (3.7) as

$$\mathrm{SNR}_i = \mathrm{SNR}_s \sum_{k\in S} |\hat{H}_{ik}|^2. \tag{3.45}$$

The same result can be proven for the cut given in Lemma 3.B.1 without requiring any special assumption. Let B denote the set of nodes located to the right of the cut but inside the rectangular slab mentioned in the lemma. See Figure 3.10. Then

$$\mathrm{SNR}_{tot} = \sum_{i\in B} \mathrm{SNR}_i + \sum_{i\in D\setminus B} \mathrm{SNR}_i. \tag{3.46}$$

For any node $i \in B$, an approximate upper bound for SNR_i is

$$\mathrm{SNR}_i \lesssim \mathrm{SNR}_s \int_0^{2\pi} \int_c^{\sqrt{n}} \frac{1}{\hat{r}^\alpha} \hat{r}\, d\hat{r}\, d\theta,$$

since Lemma 3.B.1 guarantees that there are no left-hand side nodes located at rescaled distance smaller than c to a right-hand side node i. Moreover, nodes are uniformly distributed on the network area so the summation in (3.45) over the left-hand side nodes S can be approximated by an integral. A precise upper bound on SNR_i can be found by following the binning argument used in the proof of (3.41), which yields

$$\mathrm{SNR}_i \leq K_1\,\mathrm{SNR}_s\,\log n.$$

Since there are less than $\sqrt{n}\log n$ nodes in B with high probability, the first summation in (3.46) can be upperbounded by

$$\sum_{i\in B} \mathrm{SNR}_i \leq K_1\,\mathrm{SNR}_s\,\sqrt{n}\,(\log n)^2.$$

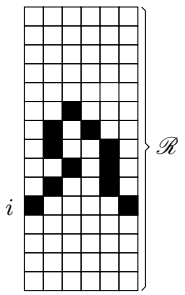

Figure 3.11: A closed left-right crossing.

Note that this contribution is smaller than any of the terms in (3.44). The second summation $\sum_{i\in D\setminus B} \mathrm{SNR}_i$ in (3.46) is equal or smaller in order to (3.44) since when the nodes B are removed there is a empty region of width at least c between the nodes S and remaining nodes $D \setminus B$. Hence for the second term in (3.46), we are back in the situation discussed in Section 3.3, hence the upperbound (3.44) applies.

Proof of Lemma 3.B.1: Let us divide our network of area $\sqrt{A} \times \sqrt{A}$ into square cells of side length $c\sqrt{A/n}$ where $0 < c < 1$ is a constant independent of A and n. We say that a cell is closed if it contains at least one node and open if it contains no nodes. Since the n nodes are uniformly and independently distributed on the network area A, the probability that a given cell is closed is upper bounded by the union bound by

$$\mathbb{P}[\text{a cell is closed}] \leq c^2.$$

Similarly, the probability that a given set of m cells $\{c_1, \ldots, c_m\}$ are simultaneously closed is upper bounded by

$$\begin{aligned}
&\mathbb{P}[\{c_1, \ldots, c_m\} \text{ is closed}] \\
&\quad = \mathbb{P}[c_1 \text{ is closed}] \times \mathbb{P}[c_2 \text{ is closed} | c_1 \text{ is closed}] \times \ldots \\
&\quad \leq c^2 \times c^2 \cdots \times c^2 \; = c^{2m}
\end{aligned} \tag{3.47}$$

since by the union bound we have,

$$\begin{aligned}
&\mathbb{P}[c_{k+1} \text{ is closed} | c_1, \ldots, c_k \text{ is closed}] \\
&\qquad\qquad \leq \frac{(c^2 A/n)}{A - k(c^2 A/n)}(n-k) \\
&\qquad\qquad \leq c^2
\end{aligned}$$

when $0 < c < 1$.

Now let us consider a slab of width $c\sqrt{A/n}\log n$ in the middle of the network. Equivalently, this is a rectangle of $\log n \times \sqrt{n}/c$ cells. By choosing c properly, we will show that this slab contains at least one *open path* that crosses

the network from top to bottom. Such a path is called an open top-bottom crossing. A path is called open if it is composed of neighboring cells that are open, a neighboring cell being one of the four cells located immediately to the top, bottom, left and right of a cell. See Fig. 3.10. On the other hand, we define a closed path in a slightly different manner: A closed path is composed of neighboring cells that are closed but a neighboring cell can now be one of the 8 cells located immediately at the top, top-left, left, bottom-left, bottom, bottom-right, right, top-right of a cell. See Fig. 3.11. With these definitions of closed and open paths, we have

$$\begin{aligned}&\mathbb{P}[\text{the slab contains an open top-bottom crossing}]\\&\quad = 1 - \mathbb{P}[\text{the slab contains a closed left-right crossing}]\end{aligned}$$

where a closed left-right crossing refers to a closed path that connects the left-boundary $\mathscr{L}$ of the slab to its right boundary $\mathscr{R}$. Let $\mathbb{P}(i \leftrightarrow \mathscr{R})$ denote the probability that there exists a closed path starting from a particular cell $i \in \mathscr{L}$ and ending at the right-boundary. Note that such a path should be at least of length $\log n$ cells. Denoting by N_i the number of closed paths of length $\log n$ that start from the cell i, we have

$$\mathbb{P}(i \leftrightarrow \mathscr{R}) \leq \mathbb{P}(N_i \geq 1).$$

By (3.47), a given path of length $\log n$ is closed with probability less than $c^{2\log n}$. By the union bound, we have

$$\mathbb{P}(N_i \geq 1) \leq c^{2\log n}\sigma_i(\log n),$$

where $\sigma_i(\log n)$ denotes the number of distinct, loop-free paths of length $\log n$ starting from i. This number is obviously not larger than $\sigma_i(\log n) \leq 5 \times 7^{(\log n - 1)}$. Combining the three inequalities, we have

$$\begin{aligned}&\mathbb{P}[\text{the slab contains a closed left-right crossing}]\\&\leq \sum_{i=1}^{\sqrt{n}/c} \mathbb{P}(i \leftrightarrow \mathscr{R}) \leq \frac{5}{7c}\sqrt{n}(7c^2)^{\log n}.\end{aligned}$$

Choosing $c^2 < \frac{1}{7\sqrt{e}}$, the last probability decreases to 0 as n increases. This concludes the proof of the lemma. □

4 Optimal Schemes

This chapter introduces four network communication schemes for wireless networks and derives their scaling performance. Each of these schemes achieves optimal scaling of the capacity in one of the four operating regimes identified in the previous chapter. The first scheme is the well-known multi-hopping scheme in the literature. We provide a brief overview of this scheme and analyze its performance in Section 4.3. We show that multi-hopping is scaling optimal only when the network is completely power-limited $\text{SNR}_s \leq 0$ dB and $\alpha \geq 3$ (Regime-III in (3.2)). The other three schemes are contributions of the current dissertation. The distributed MIMO scheme with hierarchical cooperation presented in Section 4.4 is of particular interest, since it offers a way to cope with interference in wireless networks and achieves linear scaling of the capacity when the network is not power-limited $\text{SNR}_l \geq 0$ dB (Regime-I in (3.2)). Linear scaling implies that the rate for each source-destination pair does not degrade significantly even if there are more and more users entering the system. In Section 4.5, we present a power-limited version of this scheme that is optimal when the attenuation in the network is low, $2 \leq \alpha \leq 3$ and $\text{SNR}_l < 0$ dB (Regime-II in (3.2)). A wide range of system parameters is still left out where neither of these two schemes, hierarchical cooperation nor multi-hopping, achieves optimal capacity scaling. This last regime (Regime-IV in (3.2)), has remain hidden due to the limitations of the existing scaling law formulations in the literature. In Section 4.6, we develop an optimal scheme for this regime that is a delicate combination of hierarchical cooperation and multi-hopping. Given this last hybrid scheme, the earlier two schemes, hierarchical cooperation and multi-hopping, can be viewed as the two extremes of this single unifying architecture in the respective cases when the network experiences no power-limitation $\text{SNR}_l \geq 0$ dB or when it is completely power-limited $\text{SNR}_s \leq 0$ dB.

4.1 Existing Schemes for Large Wireless Ad-Hoc Networks

As discussed in the introduction of this thesis, the literature on wireless adhoc networks can be divided into two main groups. The first group, reflecting the networking approach to wireless adhoc networks, concentrates exclusively on the multi-hopping strategy, where packets are relayed inside the network by hopping (decoding and forwarding) from one node to the next. This approach restricts the physical layer to perform simple point-to-point encoding and decoding and concentrates on the networking challenges involved. The intensive research activity in this line, especially starting from the nineties, has created a large volume of publications that lead to the establishment of numerous conferences (IEEE MASS 2009, IEEE SECON 2009, ICST AdhocNets 2009, ACM Mobihoc 2009), journals (Ad hoc Networks, Ad Hoc and Sensor Wireless Networks) and books that almost exclusively specialize on multi-hopping wireless adhoc networks.

The second line of research is rooted in network information theory. Traditionally, the aim in network information theory is to characterize the fundamental limits of performance in multi-user communication problems, usually ignoring constraints that are imposed by the current-day technology, as such constraints may not apply in general and thus are somewhat arbitrary. This approach to communication theory has been started by the seminal work [45] of C. E. Shannon in 1948 that has led to elegant characterizations and deep insights for point-to-point channels. Network information theory seeks similar characterizations for multi-user communication problems. However, multi-user communication is inherently much more complex than point-to-point communication, since even with the addition of a single helper node to the point-to-point setup, the cooperation and thus communication possibilities become diverse. Indeed, this particular setup of a network of three nodes, known as the relay channel [51, 11], remains not completely understood after almost four decades of research. Another setup that has been resistant to analysis and can serve as a basis for understanding wireless networks is the interference channel, that is, a network of four nodes, where two pairs of nodes communicate in the presence of interference from the other pair [46, 28]. The theory also has its triumphs with some multi-user settings elegantly characterized such as the multiple access channel [35, 2] and the scalar [10, 6, 7] or vector [52] Gaussian broadcast channels. Moreover, some exciting recent progress has been made in understanding the relay and interference networks, by asking coarser questions than the precise capacity region, such as degrees of freedom characterization [9] or capacity approximation within a constant number of bits [16, 4, 5].

With basic network models of three-four nodes not completely understood, network information theory fails to point out optimal strategies for large wireless adhoc networks that we consider in this thesis. However, it definitely hints the existence of alternative network communication schemes that can

potentially outperform multi-hopping. Indeed, multi-hopping corresponds to one particular form of cooperative communication, but many other forms of cooperation are known in the network information theory literature, such as amplify and forward, compress and forward, etc. See [32] and the references therein for an excellent survey on the subject. From a network information theory point of view, the contribution of the current chapter can be viewed as carefully combining various ideas on cooperation developed in this field with ideas from MIMO communication, to obtain network communication schemes whose scaling performance dramatically outperforms multi-hopping in most cases. These new strategies turn out to be more complicated than simple multi-hopping, but future wireless systems may require trading complexity for performance. The search for such schemes has been initiated by the work of Gupta and Kumar [27] in 2000. We provide below an overview of the results obtained in the literature since [27], by also discussing how they relate to the contributions of the current chapter.

The aggregate throughput scaling of the multi-hopping strategy has been characterized in [27] for the dense scaling, where the area of the network, the power budget per node and the bandwidth are kept constant as the number of users in the network increases. It has been show that the multi-hopping scheme achieves a $\sqrt{n/\log n}$ scaling of the aggregate throughput in such dense networks. Simpler derivations of the same result have been later proposed in [33], [21]. In [18], it has been shown that the $\log n$ term can be removed and that an exact scaling of $\sqrt{n}$ can be achieved by using methods from percolation theory. An aggregate throughput scaling of $\sqrt{n}$ implies that the per user rate in the network should decrease to zero as $1/\sqrt{n}$ when the number of users in the network increases. This understanding has started a new line of research that seeks for schemes that achieve better throughput scaling than multi-hopping, ideally, that scale with system size.

The first important result in this direction is the mobility scheme introduced in [26] and later extended in [14]. These works propose a relaying strategy that relies on the mobility of the users for transporting packets inside the network. The packets are carried physically from a source node to its destination by relaying nodes roaming inside the network area. Since long-distance transportation of packets is undertaken by mobility, wireless communication between nodes needs to be only of local nature between nearest-neighbor pairs. The attenuation of the signals with distance allows spatial reuse and an order of n simultaneous local communications can be established inside the network. This allows to achieve an aggregate throughput scaling linearly with the number of users. One major drawback of the scheme is that the communication delay is dictated by the velocity of the users and is therefore orders of magnitude larger than the typical delays in wireless systems. The topic of delay is further investigated in Chapter 5.

A later work [25] shows that artificial fading distributions can be constructed so that linear scaling becomes also possible in static wireless networks where the locations of the users remain fixed during the time of communica-

tion. Although the actual scheme proposed in [25] is different, it can be shown that the mobility scheme of [26] can be applied as it is, and achieves linear scaling in static networks with fading distributions from [25]. The role of mobility in [26] is now undertaken by the fading distributions of [25], that are designed carefully to allow long-distance wireless communication between two nodes without creating too much interference to the other users in the network.

In a sense, both mobility in [26] and the fading in [25] allows to circumvent the problem of interference in wireless networks. In static networks with standard fading distributions, the interference problem is central and cannot be circumvented. In order to achieve linear scaling, many simultaneous long-distance communications should be established when each of these communications constitutes interference for the others. In the point-to-point case, a physical-layer technique that achieves this is MIMO. Installing multiple antennas on both the transmitter and receiver allows to multiplex several streams and transmit them simultaneously [17, 49]. A natural way to apply this idea to networks is to group the users together to form clusters and perform MIMO transmissions between clusters. However, there is a fundamental difference between the point-to-point MIMO setup and the distributed MIMO setup in networks. In the point-to-point case, all transmit antennas and all receive antennas are located on the same device, so the streams can be jointly encoded before transmission and the received observations can be jointly decoded at the receiver. Joint processing at either the transmitter or receiver side is known to be crucial for the linear scaling of the point-to-point MIMO capacity. In the case of networks, each antenna is located on a different node and cooperation between users requires extra communication. Establishing cooperation is, therefore, the bottleneck in applying distributed MIMO ideas to wireless networks.

In [1], Aeron and Saligrama propose cooperation architecture based on coherent combining of received signals. The distributed MIMO transmissions are followed by a cooperation phase during which the nodes in the receive cluster amplify (through matched filtering) and forward their observations, so that they coherently combine at the intended destination nodes. Such coherent combining techniques have been earlier reported and analyzed in [8, 13, 36] for relay networks. The overall scheme of [1] achieves a throughput scaling of $n^{2/3}$ bits per second, as long as the received SNR in a point-to-point transmission between the farthest nodes in the network remains larger than a constant. This has been the first work to demonstrate that distributed MIMO based schemes can indeed outperform multi-hopping. The limitation to $n^{2/3}$ scaling, which is still significantly worse than the linear scaling performance of point-to-point MIMO, is precisely due to the overhead introduced by the cooperation phase.

In Section 4.4 of this chapter, we present a new multiscale, hierarchical cooperation architecture for distributed MIMO transmission that does not introduce significant overhead to communication. The key observation behind the hierarchical architecture is that cooperation is itself another communication problem. This observation allows to use any known scheme for commu-

nicating in wireless networks to establish cooperation for distributed MIMO communication, which overall yields a better communication scheme for wireless networks than the scheme we begin with. Applying this idea recursively builds a hierarchical architecture that achieves an aggregate throughput scaling $n^{1-\epsilon}$ for any $\epsilon > 0$ in networks with $\mathrm{SNR}_l \geq 0$ dB. Note that the condition $\mathrm{SNR}_l \geq 0$ dB is weaker than requiring a constant SNR in a point-to-point transmission between farthest nodes by a factor of $1/n$. Recall that SNR_l was defined in (2.5) as n times the SNR between farthest nodes. The power gain comes from the MIMO effect.

When the network experiences a power limitation ($\mathrm{SNR}_l < 0$ dB), distributed MIMO communication can still be beneficial but the aggregate throughput can not scale with system size anymore. Distributed MIMO based schemes still outperform multi-hopping, unless the power limitation in the network is so severe that nodes cannot communicate efficiently beyond their nearest neighbors ($\mathrm{SNR}_s < 0$ dB). Nevertheless, hierarchical cooperation outperforms multi-hopping when $2 \leq \alpha < 3$ even if $\mathrm{SNR}_s < 0$ dB. In the power-limited regime, distributed MIMO communication not only provides a degrees of freedom gain, but also provides a power gain obtained by combining signals received at different nodes. However, the distributed MIMO scheme with hierarchical cooperation in Section 4.4 cannot be applied as it is to such power-limited networks. To achieve optimal scaling, two different modifications of the scheme are presented in Section 4.5 and Section 4.6 for the respective cases where the power path loss attenuation in the network is low ($2 \leq \alpha < 3$) or high ($\alpha \geq 3$). The scheme presented in Section 4.6 is of particular interest, as it delicately combines distributed MIMO communication with multi-hopping.

4.2 Main Result

The three main results of this chapter are summarized in the following three theorems. The theorems characterize the performance achieved by the three new schemes presented in this chapter. We consider the model introduced in Section 2, except that we state the achievability results below in a slightly more general setting. The results apply to any source-destination pairing in the network. For example, each source node can be associated to the destination node located farthest or nearest to itself. They also apply to pairings that are obtained by random association of source nodes with destination nodes, i.e., the random pairing introduced in Section 2. Note that even if the source-destination pairing is arbitrary, the node locations are not; we still assume that the n nodes are distributed uniformly and independently on the network area.

Theorem 4.2.1. *Let $\mathrm{SNR}_l \geq 0$ dB for all n. For any $\epsilon > 0$, there exists a constant $K_\epsilon > 0$ independent of n such that w.h.p an aggregate throughput*

$$T \geq K_\epsilon \, n^{1-\epsilon}$$

is achievable in a random realization of the network for all possible pairings between sources and destinations.

This performance is achieved by distributed MIMO communication with hierarchical cooperation and the theorem is proven in Section 4.4. Recall from Definition 2.2.2 that a scheme and its throughput are defined in the scaling sense.

Theorem 4.2.2. *Let $SNR_l < 0$ dB for all n. For any $\epsilon > 0$, there exists a constant $K_\epsilon > 0$ independent of n such that w.h.p an aggregate throughput*

$$T \geq K_\epsilon \, n^{1-\epsilon} SNR_l$$

is achievable in a random realization of the network for all possible pairings between sources and destinations.

Note that $\mathrm{SNR}_l < 0$ dB, or $\mathrm{SNR}_l = O(1)$. Hence, the aggregate throughput in Theorem 4.2.2 does not scale linearly with the number of users. The exact scaling depends on how fast SNR_l decreases to 0 with increasing n. This performance is achieved by a modification of the distributed MIMO communication scheme with hierarchical cooperation. The theorem is proven in Section 4.5.

Theorem 4.2.3. *Let $\alpha > 2$, $SNR_s > 0$ dB and $SNR_l < 0$ dB for all n.*[1] *For any $\epsilon > 0$, there exists a constant $K_\epsilon > 0$ independent of n such that w.h.p an aggregate throughput*

$$T \geq K_\epsilon \sqrt{n} \, SNR_s^{\frac{1}{\alpha-2}-\epsilon}, \tag{4.1}$$

is achievable in a random realization of the network for all possible pairings between sources and destinations.

Note that since $\mathrm{SNR}_s > 0$ dB, or $\mathrm{SNR}_s = \Omega(1)$, the aggregate throughput in Theorem 4.2.3 is $o(\sqrt{n})$. On the other hand, recall that

$$\mathrm{SNR}_s = n^{\alpha/2-1}\mathrm{SNR}_l,$$

and $\mathrm{SNR}_l = O(1)$, so it can be verified that the aggregate throughput in (4.1) is $O(n)$. Therefore, the aggregate throughput scaling achieved in Theorem 4.2.3 is between $\sqrt{n}$ and n and the exact scaling is dictated by the scaling of SNR_s. This performance is achieved by a hybrid architecture combining distributed MIMO communication with multi-hopping. The theorem is proven in Section 4.6.

Combining the results of the three theorems with the scaling achieved by multi-hopping in (4.3) yields the following lower bound on the scaling exponent

[1] $\alpha = 2$ is excluded as $\mathrm{SNR}_s = \mathrm{SNR}_l$ in that case.

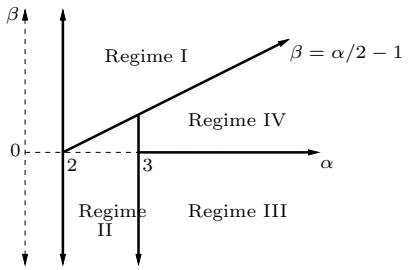

Figure 4.1: The four operating regimes. The optimal schemes in these regimes are I-Hierarchical Cooperation, II-Bursty Hierarchical Cooperation, III-Multihop, IV- Multihop MIMO Hierarchical Cooperation.

of the aggregate throughput,

$$e(\alpha,\beta) \geq \begin{cases} 1 & \beta \geq \alpha/2-1 & \text{Hierarchical Cooperation} \\ 2-\alpha/2+\beta & \beta < \alpha/2-1 & \text{Power-Limited HC} \\ & \text{and } 2 \leq \alpha < 3 & \\ 1/2+\beta & \beta \leq 0 & \text{Multihopping} \\ & \text{and } \alpha \geq 3 & \\ 1/2+\beta/(\alpha-2) & 0 < \beta < \alpha/2-1 & \text{Multihopping with HC,} \\ & \text{and } \alpha \geq 3 & \end{cases} \tag{4.2}$$

for any real β, where β is the scaling exponent of SNR_s defined earlier in (2.9). Together with (3.2), this result establishes the capacity scaling of wireless networks. The four regimes in (4.2) are shown in Figure 4.1.

4.3 Nearest-Neighbor Multihopping

We start by briefly recalling the basic properties of the multi-hopping scheme and deriving its scaling performance. In the light of the new formulation introduced in this thesis, we will see that the nearest neighbor multi-hopping scheme achieves a $\sqrt{n}$ scaling of the aggregate throughput only when the nearest neighbor $\mathrm{SNR}_s \geq 0$ dB. When $\mathrm{SNR}_s < 0$ dB, it achieves a throughput scaling of $\sqrt{n}\,\mathrm{SNR}_s$. The result can be rewritten in terms of the scaling exponent of the aggregate throughput as

$$e_{multihop}(\alpha,\beta) = \begin{cases} 1/2 & \beta \geq 0 \\ 1/2+\beta & \beta < 0, \end{cases} \tag{4.3}$$

where β is the scaling exponent of SNR_s.

Let us start by dividing the network into square cells of area $A_c = 2A\log n/n$. According to Lemma 2.3.1, each cell contains at least one node w.h.p in a random realization of the network. In the multi-hopping scheme, the messages

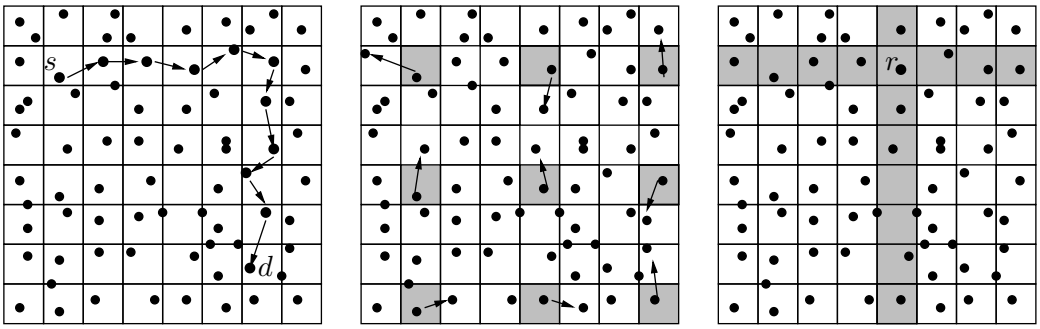

Figure 4.2: The Multihopping Scheme: a) The packets of a source node s are delivered to its destination node d by multi-hopping from one cell to the next. b) A 9-TDMA scheme is employed between cells to control inter-cluster interference. The shaded cells are simultaneously active. c) The relaying traffic at node r is originated from or destined to one of the nodes located in the shaded rectangles.

are relayed from source nodes to their destinations by hopping from one cell to the next. See Figure 4.2-(a). The fact that all cells are non-empty ensures that we can find a node in each cell to assign the relaying job. This relay node decodes the messages transmitted from its neighboring cells, temporarily stores, re-encodes and forwards them to the next cell in their respective direction of transportation. Hence, the communication in the network is based on point-to-point transmissions between pairs of nodes located in neighboring cells. Most of these transmissions are from relay-to-relay nodes but we also have source-to-relay and relay-to-destination node transmissions in the respective cases when a source node drains or a destination node collects its messages. The signal-to-noise power ratio in these point-to-point transmissions is lower bounded by

$$\mathrm{SNR}_{relay} \geq \frac{\mathrm{SNR}_s}{(10 \log n)^{\alpha/2}},$$

since the separation between the transmitting and the receiving node is upper bounded by $\sqrt{5A_c}$. In order to be able to control interference, not all cells are allowed to operate simultaneously. Instead, a TDMA strategy is employed between cells so that only a constant fraction of the cells are active in a given time-slot and there are a number of inactive cells between every pair of active cells. Nodes in a cell are allowed to transmit only when their cell is active according to the TDMA scheme. Otherwise they remain silent. The mid-figure in Figure 4.2 illustrates a 9-TDMA strategy. The interference-to-noise-power ratio received by any node in the network from the simultaneously active clusters according to the 9-TDMA scheme, is upper bounded by

$$\mathrm{INR}_{relay} \leq K_I \, \mathrm{SNR}(A_c) = K_I \, \frac{\mathrm{SNR}_s}{(2 \log n)^{\alpha/2-1}}$$

for a constant K_I independent of n and SNR_s. (This result can be shown by a simple modification of Lemma 4.4.3 in the special case $A_c = 2A\log n/n$. The lemma is proven later in Section 4.4.1.) Therefore, the rate achieved in a point-to-point transmission between two neighboring cells by treating all the inter-cell interference as noise is lower bounded by

$$R_{relay} = \log\left(1 + \frac{\mathrm{SNR}_{relay}}{1+\mathrm{INR}_{relay}}\right) \geq \log\left(1 + \frac{\mathrm{SNR}_s(10\log n)^{-\alpha/2}}{1+K_I\mathrm{SNR}_s(2\log n)^{1-\alpha/2}}\right). \tag{4.4}$$

Note that the actual relaying rate is $R_{relay}/9$ since each relay node gets to transmit once in 9 time slots according to the 9-TDMA scheme. The draining rate $R_{relay}/9$ of a given relay node has to be shared between the source-destination pairs whose path is routed through this relay node. We assume a simplistic route between the source-destination pairs where packets are first relayed through a horizontal slab containing the source node and then inside a vertical slab containing the destination node. See Figure 4.2-(a). These horizontal and vertical slabs of area $\sqrt{A_c} \times \sqrt{A}$ contain less than $\sqrt{A_cA} \times \frac{n}{A} = \sqrt{n\log n}$ nodes w.h.p. (This can be seen by applying part-(e) of Lemma 2.3.1 to the horizontal and vertical slabs.) With this routing strategy, the relaying traffic generated at a relay node is due to source nodes located in the same horizontal slab or destination nodes located in the same vertical slab. See Figure 4.2-(c). Hence, the relaying rate of (4.4) has to be shared among at most $2\sqrt{n\log n}$ source-destination pairs. This leads to an overall rate

$$R_{multihop} \geq \frac{1}{18\sqrt{n\log n}} R_{relay}. \tag{4.5}$$

per source-destination pair. The aggregate throughput achieved by multihopping is the lower-bounded by

$$T_{multihop} \geq K_0\sqrt{\frac{n}{\log n}}\log\left(1 + (\log n)^{-\alpha/2}\frac{\mathrm{SNR}_s}{1+K_I\mathrm{SNR}_s}\right)$$

for a constant $K_0 > 0$ independent of n, which yields the scaling exponents in (4.3). Using percolation theory, the above argument can be refined to yield an aggregate throughput $T_{multihop} \geq K_0\sqrt{n}\log\left(1+\frac{\mathrm{SNR}_s}{1+K_I\mathrm{SNR}_s}\right)$. Note that in the above derivation, we have not made any assumption on the source-destination pairing, so the lower bound (4.5) holds for all possible pairings of the source nodes with destination nodes.

4.4 Distributed MIMO with Hierarchical Cooperation

In this section, by proving Theorem 4.2.1, we will show that linear scaling is achievable in wireless networks when $\mathrm{SNR}_l \geq 0$ dB. Note that the traditional

multi-hopping scheme achieves only an aggregate throughput scaling of $\sqrt{n}$ in this case. The proof of Theorem 4.2.1 relies on the construction of an explicit scheme that realizes the promised scaling law. The construction is based on recursively using the following key lemma, which addresses the case when $\alpha > 2$ and whose proof is relegated to Section 4.4.1.

Lemma 4.4.1. *Consider $\alpha > 2$ and the network with n nodes subject to interference from external sources. The signal received by node i is given by*

$$Y_i = \sum_{k \neq i} H_{ik} X_k + Z_i + I_i$$

where I_i is the external interference signal received by node i. Let the long distance SNR in the network be lower bounded,

$$SNR_l \geq K_S \tag{4.6}$$

for a constant $K_S > 0$ independent of n. Assume also that $\{I_i, 1 \leq i \leq n\}$ is a collection of uncorrelated zero-mean stationary and ergodic random processes with interference to noise power ratio upper bounded by

$$INR_{ex} \leq K_I \, SNR_l \tag{4.7}$$

for a constant $K_I > 0$ independent of n and SNR_l. Let us assume that there exists a scheme such that for each n, with probability at least $1 - e^{-n^{c_1}}$, it achieves an aggregate throughput

$$T \geq K_1 \, n^b$$

for all possible source-destination pairings in a random realization of the network. K_1 and c_1 are positive constants independent of n and the source-destination pairing, and $0 \leq b < 1$.

Then one can construct another scheme for this network that achieves a higher *aggregate throughput scaling*

$$T \geq K_2 \, n^{\frac{1}{2-b}}$$

for all possible source-destination pairings, where $K_2 > 0$ is another constant independent of n and the pairing. Moreover, the failure probability for the new scheme is upper bounded by $e^{-n^{c_2}}$ for another positive constant c_2.

Lemma 4.4.1 is the key step to build a hierarchical architecture. Since $\frac{1}{2-b} > b$ for $0 \leq b < 1$, the new scheme is always better than the old one. We will now give a rough description of how the new scheme can be constructed given the old scheme, as well as a back-of-the-envelope analysis of the scaling law it achieves. Next section is devoted to the precise description and performance analysis of the scheme.

The scheme that proves Lemma 4.4.1 is based on clustering and long-range MIMO transmissions between clusters. We divide the network into clusters of M nodes. Let us focus for now on a particular source node s and its destination node d. s sends M bits to d in three steps:

(1) Node s distributes its M bits among the M nodes in its cluster, one for each node;

(2) These nodes together can then form a distributed transmit antenna array, sending the M bits *simultaneously* to the destination cluster where d lies;

(3) Each node in the destination cluster gets one observation from the MIMO transmission and it quantizes and ships the observation to d, which can then do joint MIMO processing of all the observations and decode the M transmitted bits.

From the network point of view, all source-destination pairs have to eventually accomplish these three steps. Step 2 is long-range communication and only one source-destination pair can operate at at a time. Steps 1 and 3 involve local communication and can be parallelized across source-destination pairs. Combining all this leads to three phases in the operation of the network:

Phase 1: Setting Up Transmit Cooperation Clusters work in parallel. Within a cluster, each source node has to distribute M bits to the other nodes, 1 bit for each node, such that at the end of the phase, each node has 1 bit from each of the source nodes in the same cluster. Since there are M source nodes in each cluster, this gives a total traffic of exchanging $M(M-1) \sim M^2$ bits. (Recall our assumption that each node is a source for some communication request and a destination for another.) The key observation is that this is similar to the original problem of communicating between n source and destination pairs, but on a smaller network of size M. More precisely, this traffic demand of exchanging M^2 bits can be handled by setting up M sub-phases, and assigning M source-destination pairs for each sub-phase to communicate their 1 bit. Since our channel model is scale invariant, the scheme given in the hypothesis of the lemma can be used in each sub-phase. With a scheme achieving aggregate throughput M^b, each sub-phase is completed in $M^{1-b}/2$ time slots, so the whole phase takes M^{2-b} time slots. See Figure 4.3.

Phase 2: MIMO Transmissions We perform successive long-distance MIMO transmissions between source-destination pairs, one at a time. In each one of the MIMO transmissions, say one between s and d, the M bits of s are simultaneously transmitted by the M nodes in its cluster to the M nodes in the cluster of d. Each of the long-distance MIMO transmissions are repeated for each source-destination pair in the network, hence we need n time slots to complete the phase. See Figure 4.4.

Phase 3: Cooperate to Decode Clusters work in parallel. Since there are M destination nodes inside each cluster, each cluster has received M MIMO

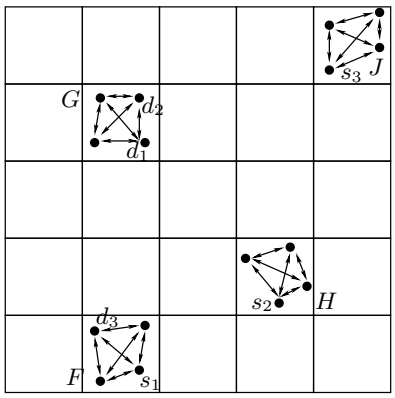

Figure 4.3: Nodes inside clusters F, G, H and J are illustrated while exchanging bits in Phases 1 and 3. Note that in Phase 1 the exchanged bits are the source bits whereas in Phase 3 they are the quantized MIMO observations. Clusters work in parallel. In this and the following figure Fig. 4.4, we highlight three source-destination pairs $s_1 - d_1$, $s_2 - d_2$ and $s_3 - d_3$, such that nodes s_1 and d_3 are located in cluster F, nodes s_2 and s_3 are located in clusters H and J respectively, and nodes d_1 and d_2 are located in cluster G.

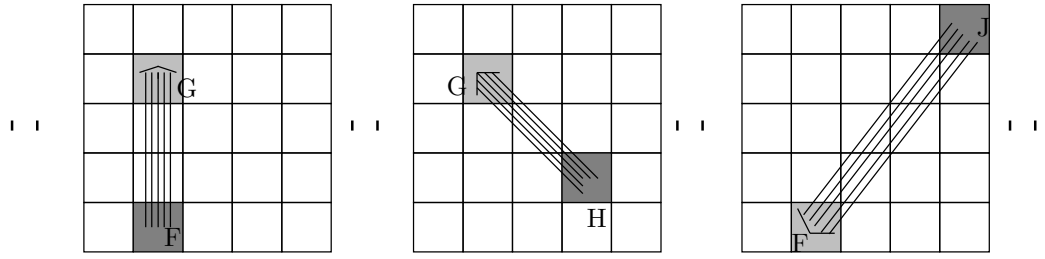

Figure 4.4: Successive MIMO transmissions are performed between clusters. The first figure depicts MIMO transmission from cluster F to G, where bits originally belonging to s_1 are simultaneously transmitted by all nodes in F to all nodes in G. The second MIMO transmission is from H to G, while now bits of source node s_2 are transmitted by nodes in H to nodes in G. The third picture illustrates MIMO transmission from cluster J to F.

transmissions in phase 2. Each MIMO transmission is intended for a different destination node. Thus, each node in the cluster has M received observations, one from each of the MIMO transmissions, and each observation is to be conveyed to a different destination node in its cluster. Nodes quantize each observation into fixed Q bits, so there are now a total of QM^2 bits to exchange inside each cluster. Using exactly the same scheme as in Phase 1, we conclude the phase in QM^{2-b} time slots. See again Figure 4.3.

Assuming that each destination node is able to decode the transmitted bits from its source node from the M quantized signals it gathers by the end of Phase 3, we can calculate the rate of the scheme as follows. Each source node is able to transmit M bits to its destination node, hence nM bits in total are delivered to their destinations in $M^{2-b} + n + QM^{2-b}$ time slots, yielding an aggregate throughput of

$$\frac{nM}{M^{2-b} + n + QM^{2-b}}$$

bits per time slot. Maximizing this throughput by choosing $M = n^{\frac{1}{2-b}}$ yields $T(n) = \frac{1}{2+Q} n^{\frac{1}{2-b}}$ for the aggregate throughput, which is the result in Lemma 4.4.1.

Clusters can work in parallel in phases 1 and 3 because for $\alpha > 2$, the inter-cluster interference power to noise ratio experienced by any node in the network is bounded by a constant fraction of the long-distance SNR in the cluster. In other words, the condition (4.7) is satisfied simultaneously for all clusters in the network. Moreover, the interference signals received by different nodes in the cluster are zero-mean and uncorrelated, satisfying therefore the assumptions of Lemma 4.4.1. For $\alpha = 2$, the inter-cluster interference to noise power ratio scales like $\log n$ times the long distance SNR in the cluster, leading to a slightly different version of Lemma 4.4.1, whose proof is given at the end of Section 4.4.1.

Lemma 4.4.2. *Consider $\alpha = 2$ and the network with n nodes subject to interference from external sources. The signal received by node i is given by*

$$Y_i = \sum_{k \neq i} H_{ik} X_k + Z_i + I_i$$

where I_i is the external interference signal received by node i. Let the long distance SNR in the network be lower bounded,

$$SNR_l \geq K_S$$

for a constant $K_S > 0$ and independent of n. Assume also that $\{I_i, 1 \leq i \leq n\}$ is a collection of uncorrelated zero-mean stationary and ergodic random processes with interference to noise power ratio upper bounded by

$$INR_{ex} \leq K_I \, SNR_l \, \log n,$$

for a constant $K_I > 0$ independent of n and SNR_l. Let us assume there exists a scheme such that for each n with failure probability at most $e^{-n^{c_1}}$, achieves an aggregate throughput

$$T \geq K_1 \frac{n^{b_1}}{(\log n)^{b_2}}$$

for all possible source-destination pairings in the network, where K_1 and c_1 are positive constants independent of n and the source-destination pairing, and $0 \leq b_1 < 1$, $b_2 \geq 0$.

Then, one can construct another scheme for this network that achieves a higher *aggregate throughput scaling*

$$T \geq K_2 \frac{n^{\frac{1}{2-b_1}}}{(\log n)^{b_2+1}}$$

for every possible source-destination pairing in the network, where $K_2 > 0$ is another constant independent of n and the pairing. Moreover, the failure probability for the new scheme is upper bounded by $e^{-n^{c_2}}$ for another positive constant c_2.

We can now use Lemma 4.4.1 and 4.4.2 to prove Theorem 4.2.1.

Proof of Theorem 4.2.1: We only focus on the case $\alpha > 2$ and comment on the extension to the case $\alpha = 2$ at the end of the proof.

To prove Theorem 4.2.1 for the case $\alpha > 2$, we consider Lemma 4.4.1. We start by observing that the simple scheme of transmitting directly from source nodes to their destination nodes in a round-robin fashion (TDMA) achieves an aggregate throughput $\Theta(1)$ under the conditions of the lemma. If $\mathrm{SNR}_l \geq K_S$, the received SNR in a point-to-point transmission between any source-destination pair in the network is lowerbounded by $\mathrm{SNR}_l/n \geq K_S/n$. Note that the lower-bound SNR_l/n corresponds to the case where the source-destination pair is separated by the largest distance, the diameter of the network. In the TDMA scheme, source nodes transmit only a fraction $\frac{1}{n}$ of the time and remain inactive otherwise. Therefore, when active, they can transmit with power Pn instead of P and still satisfy their average power constraint P. This results in received SNR equal to $n \times \frac{\mathrm{SNR}_l}{n} \geq K_S$ during transmission. Since the external INR experienced by destination nodes is bounded above by $K_I \times \mathrm{SNR}_l$, each source-destination pair can communicate at a rate, bounded below by

$$R_{TDMA} \geq \frac{1}{n} \log\left(1 + \frac{\mathrm{SNR}_l}{1 + K_I \mathrm{SNR}_l}\right) \geq \frac{1}{n} \log\left(1 + \frac{K_S}{1 + K_I K_S}\right), \qquad (4.8)$$

yielding an aggregate throughput $\Theta(1)$, or $b = 0$. The failure probability is 0, since the strategy can be operated in any realization of the random network; for arbitrary placement of nodes, arbitrary source-destination pairings and arbitrary channel states.

As soon as we have a scheme to start with, Lemma 4.4.1 can be applied recursively, yielding a scheme that achieves higher throughput at each step of the recursion. More precisely, starting with a TDMA scheme corresponding to $b = 0$ and applying Lemma 4.4.1 recursively h times, one gets a scheme achieving $\Theta(n^{\frac{h}{h+1}})$ aggregate throughput. Given any $\epsilon > 0$, we can now choose h such that $\frac{h}{h+1} \geq 1 - \epsilon$ and we get a scheme that achieves $\Theta(n^{1-\epsilon})$ aggregate throughput scaling with high probability.

To prove Theorem 4.2.1 for the case $\alpha = 2$, we consider Lemma 4.4.2. The TDMA strategy achieves an aggregate rate $\Theta(1/\log n)$, corresponding to $b_1 = 0$ and $b_2 = 1$, under the external interference given in the lemma. The lemma can be applied recursively h times, and yields a scheme with aggregate throughput $\Theta(n^{\frac{h}{h+1}}/(\log n)^{h+1})$. Given any $\epsilon > 0$, we can now choose h such that $\frac{h}{h+1} > 1-\epsilon$ and we get a scheme that achieves $\Theta(n^{1-\epsilon})$ aggregate throughput scaling with high probability. This concludes the proof of Theorem 4.2.1. □

Remark 4.4.1. *The recursive application of Lemma 4.4.1 actually implies a stronger result than the one stated in Theorem 4.2.1. The aggregate throughput scaling $\Theta(n^{1-\epsilon})$ is also achievable when the network experiences external interference satisfying the conditions of Lemma 4.4.1.*

Gathering everything together, we have built a hierarchical scheme to achieve the desired throughput. At the lowest level of the hierarchy, we use the simple TDMA scheme to exchange bits for cooperation among small clusters. Combining this with longer range MIMO transmissions, we get a higher throughput scheme for cooperation among nodes in larger clusters at the next level of the hierarchy. Finally, at the top level of the hierarchy, the cooperation clusters are almost the size of the network and the MIMO transmissions are over the global scale to meet the desired traffic demands. Figure 4.5 shows the resulting hierarchical scheme, with a focus on the top two levels.

4.4.1 Detailed Description and Performance Analysis

In this section, we concentrate in more detail on the scheme that proves Lemma 4.4.1 and Lemma 4.4.2. We first focus on Lemma 4.4.1 and then extend the proof to Lemma 4.4.2. As we have already seen in the previous section, we start by dividing the network area A into smaller squares of area $A_c = M\frac{A}{n}$. Since the node density is n/A, there will be on average M nodes inside each of these small squares. Lemma 2.3.1-(e) upperbounds the probability of having large deviations from the average. Applying Lemma 2.3.1-(e) to the squares of area $M\frac{A}{n}$, we see that all squares contain order M nodes with probability larger than $1 - \frac{n}{M}e^{-\Lambda(\delta)M}$. In the sequel, we assume $M = n^{\mu}$, for a constant $0 < \mu \leq 1$, in which case this probability tends to 1 as n increases. This condition is sufficient for our below analysis on scaling laws to

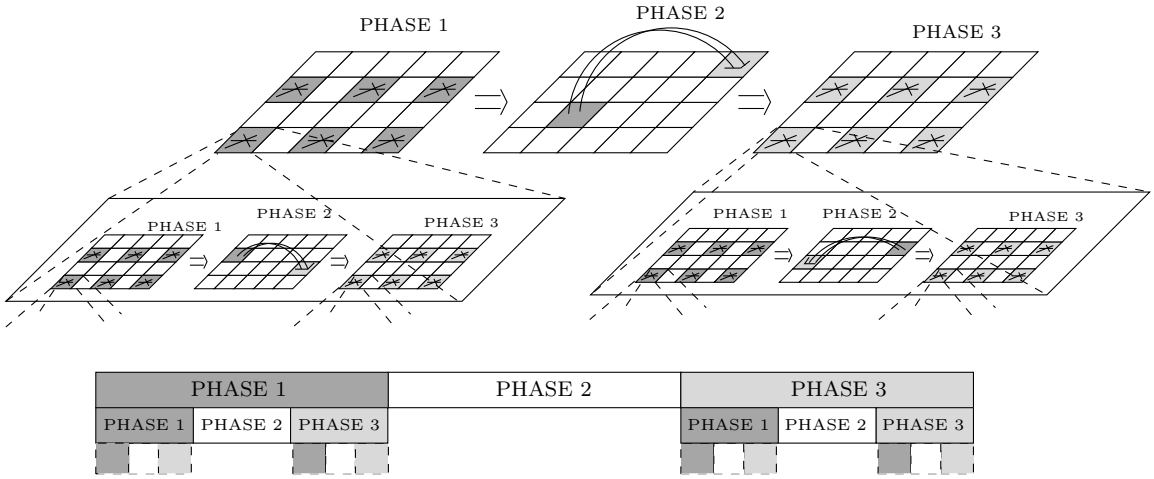

Figure 4.5: The upper figure illustrates the salient features of the three phase hierarchical scheme. The time division in this hierarchical scheme is explicitly given the figure below.

hold. However, in order to simplify the presentation, we will assume that there are exactly M nodes in each square.

The clustering is used to realize a distributed MIMO system in three successive steps:

Phase 1: Setting Up Transmit Cooperation In this phase, source nodes distribute their data streams over their clusters and set up the stage for the long-range MIMO transmissions that we want to perform in the next phase. Clusters work in parallel according to the 9-TDMA scheme depicted in Figure 4.6, which divides the total time for this phase into 9 time-slots and assigns simultaneous operation to clusters that are sufficiently separated. The shaded clusters in Figure 4.6 are operating simultaneously in the same time slot while the other clusters stay inactive. Note that with this scheduling, in every time slot there are at least two inactive clusters between any two clusters that are active.

Let us focus on one specific source node s located in cluster S with destination node d in cluster D. Node s will divide a block of length LM bits of its data stream into M sub-blocks each of length L bits, where L can be arbitrarily large and can depend on M or n. The destination of each sub-block in Phase 1 depends on the relative positions of clusters S and D:

(1) If S and D are either the same cluster or are not neighboring clusters: one sub-block is to be kept in s and the remaining $M-1$ sub-blocks are to be distributed among the other $M-1$ nodes located in S, one sub-block for each node.

(2) If S and D are neighboring clusters: Divide the cluster S into two halves, each of area $A_c/2$, one half located close to the border with D and the

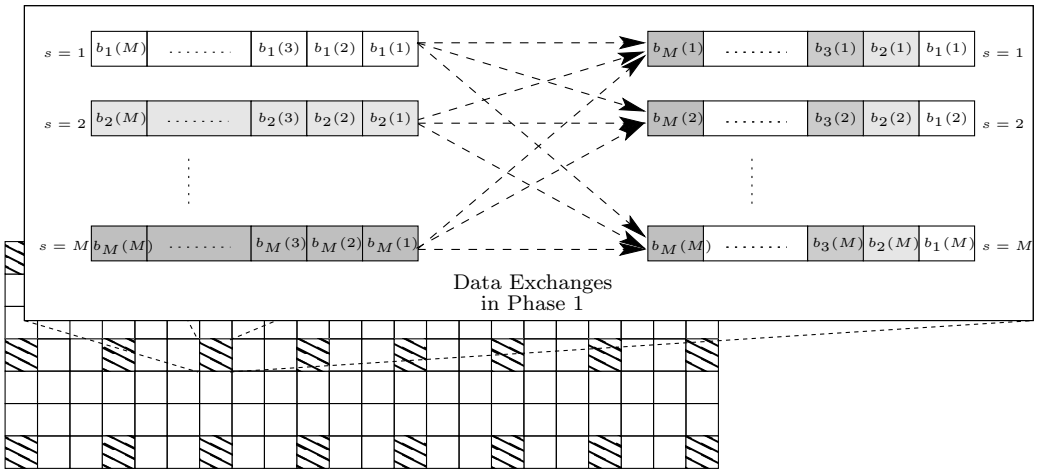

Figure 4.6: Buffers of the nodes in a cluster are illustrated before and after the data exchanges in Phase 1. The data stream of the source nodes are distributed to the M nodes in the network as depicted. $b_s(j)$ denotes the j'th sub-block of the source node s. Note the 9-TDMA scheme that is employed over the network in this phase.

second half located farther to D. The M sub-blocks of source node s are to be distributed to the $M/2$ nodes located in the second half cluster (farther to D), each node gets two sub-blocks. The fact that each half cluster contains $\Theta(M/2)$ nodes w.h.p. can be concluded from Lemma 2.3.1-(e).

The above traffic is required for every node in cluster S since every node is a source for some communication request. Thus, each node in S needs to distribute its data among the rest $M-1$ nodes in the cluster, which gives a highly uniform traffic demand of delivering $M \times L(M-1) \sim LM^2$ bits in total to their destinations. A key observation is that the problem can be separated into sub-problems, each similar to our original problem, but on a network of size M and area A_c. More specifically, the traffic of transporting LM^2 bits can be handled by organizing M sessions and assigning M source-destination pairs for each session. (Note that due to the non-uniformity arising from point (2) above, one might be able to assign only $M/2$ source-destination pairs in some sessions and hence need to handle the traffic demand of transporting LM^2 bits by organizing up to $2M$ sessions instead of M.) The assigned source-destination pairs in each session can then communicate their sub-block of L bits. Since our channel model is scale invariant, the scheme assumed in the hypothesis of Lemma 4.4.1 can be used to handle the traffic in each session. However, we first need to verify that conditions (4.6) and (4.7) of the lemma are satisfied for each cluster. Note that the long-distance SNR in each cluster,

$\mathrm{SNR}(A_c)$, is larger than the long-distance SNR in the network,

$$\mathrm{SNR}(A_c) > \mathrm{SNR}_l \geq K_S.$$

Hence, condition (4.6) is satisfied. The scheme is to be operated simultaneously inside all the clusters in the 9-TDMA scheme, so we need to ensure that the resultant inter-cluster interference satisfies the properties in Lemma 4.4.1.

Lemma 4.4.3. *Consider clusters operating simultaneously according to the 9-TDMA scheme in Figure 4.6. Let the long-distance SNR in each cluster be SNR(A_c). For $\alpha > 2$, the interference experienced by each node from the simultaneously operating clusters satisfies*

$$INR_{ic} \leq K_{I_1} SNR(A_c),$$

where K_{I_1} is constant independent of n and SNR(A_c). When $\alpha = 2$, the inter-cluster interference to noise power is bounded as

$$INR_{ic} \leq K_{I_2} \log n \, SNR(A_c),$$

for a constant K_{I_2} independent of n and SNR(A_c). Moreover, the interference signals received by different nodes in the cluster are zero-mean and uncorrelated.

The proof of this lemma is given at the end of the present section. Let us for now concentrate on the case $\alpha > 2$. By Lemma 4.4.3, the inter-cluster interference to noise power ratio is bounded by a constant times the long-distance SNR in the cluster and the interference signals experienced by different nodes are uncorrelated. Hence, the strategy in the hypothesis of Lemma 4.4.1 can achieve an aggregate rate $K_1 M^b$ in each session for some constant $K_1 > 0$, with probability larger than $1 - e^{-M^{c_1}}$. Using the union bound, with probability larger than $1 - 2ne^{-M^{c_1}}$, the aggregate rate $K_1 M^b$ is achieved inside all sessions in all clusters in the network. (Recall that the number of sessions in one cluster can be $2M$ in the extreme case and there are n/M clusters in total.) With this aggregate rate, each session can be completed in at most $(L/K_1)M^{1-b}$ channel uses and $2M$ successive sessions are completed in $(2L/K_1)M^{2-b}$ channel uses. Using the 9-TDMA scheme, the phase is completed in less than $(18L/K_1)M^{2-b}$ channel uses all over the network with probability larger than $1 - 2ne^{-M^{c_1}}$.

Phase 2: MIMO Transmissions In this phase, we are performing successive long-distance MIMO transmissions between clusters. A MIMO transmission from source node s to destination node d involves the M (or $M/2$) nodes in the cluster S of s, and the M (or $M/2$) nodes in the cluster D of d. The cluster S is referred as the source cluster for this MIMO transmission and D is the destination cluster.

If S and D are the same cluster, we skip the step for this source-destination pair $s-d$. Otherwise, we operate in two slightly different modes depending on the relative positions of S and D. Each mode is a continuation of the operations performed in the first phase. First consider the case where S and D are not neighboring clusters. In this case, the M nodes in cluster S independently encode the L bits-long sub-blocks they possess, originally belonging to node s, into C channel symbols by using a randomly generated Gaussian code $\mathcal{C}$. The nodes then transmit their encoded sequences of length C symbols simultaneously to the M nodes in cluster D. The nodes in cluster D quantize the signals they observe during the C transmissions and store these quantized signals(that we will simply refer to as *observations* in the following text), without trying to decode the transmitted symbols. In the case where S and D are neighbors, the strategy is slightly modified so that the MIMO transmission is from the $M/2$ nodes in S, that possess the sub-blocks of s after Phase 1, to the $M/2$ nodes in D that are located in the farther half of the cluster to S. Each of these $M/2$ nodes in S possess two sub-blocks that come from s. They encode each sub-block into C symbols by again using a Gaussian codebook. The nodes then transmit the $2C$ symbols to the $M/2$ nodes in D that in turn sample their received signals and store the observations. The observations accumulated at various nodes in D at the end of this step are to be conveyed to node d during the third phase.

After concluding the step for the pair $s-d$, the phase continues by repeating the same step for the next source node $s+1$ in S and its destination d'. Note that the destination cluster for this new MIMO transmission is, in general, a different cluster D', which is the one that contains the destination node d'. The MIMO transmissions are repeated until the data originated from all source nodes in the network are transmitted to their respective destination clusters. Since the step for one source-destination pair takes either C or $2C$ channel uses, completing the operation for all n source nodes in the network requires at most $2C \times n = 2Cn$ channel uses.

Note that we perform n MIMO transmissions in total, one for each source-destination pair in the network. On the other hand, there are M source nodes in each cluster so each node in the network is transmitting only during M of these MIMO transmissions. In other words, each node is active only during a fraction M/n of the total duration of the phase. Therefore, the nodes can use a Gaussian codebook with elevated power nP/M and still satisfy their average power constraint P. We make use of this fact while proving Lemma 4.4.4 that is given in the sequel.

Phase 3: Cooperate to Decode In this phase, we aim to provide each destination node, the observations of the symbols that have been originally intended for it. With the MIMO transmissions in the second phase, these observations are accumulated at the nodes of its cluster. As before, let us focus on a specific destination node d located in cluster D. Note that depending on whether the source node of d is located in a neighboring cluster or not, either each of the M nodes in D have C observations intended for d, or $M/2$ of

the nodes have $2C$ observations each. Note that these observations are some real numbers that need to be quantized and encoded into bits before being transmitted. Let us assume that we are encoding each block of C observations into CQ bits, by using fixed Q bits per observation on the average. The situation is symmetric for all nodes in D since they are all destinations for some traffic. Hence, the cluster has received M MIMO transmissions in the previous phase, one for each destination node. (The destination nodes that have source nodes in D are exception. Recall from Phase 1 and Phase 2 that in this case, each node in D possesses sub-blocks of the original data stream for the destination node and not MIMO observations. We will ignore this case by simply assuming $L \leq CQ$ in the below computation.) The resulting traffic demand of transporting $M \times CQM$ bits in total is similar to Phase 1 and can be handled by using exactly the same scheme in less than $(2CQ/K_1)M^{2-b}$ channel uses. Recalling the discussion on the first phase, we conclude that the phase can be completed in less than $(18CQ/K_1)M^{2-b}$ channel uses all over the network with probability larger than $1 - 2ne^{-M^{c_1}}$.

Note that if it were possible to encode each observation into fixed Q bits without introducing any distortion, which is obviously not the case, the following lemma on MIMO capacity, would imply that with the Gaussian code $\mathcal{C}$ used in Phase 2 satisfying $L/C \geq \kappa$ for some constant $\kappa > 0$ independent of M and n, the transmitted bits could be recovered by an arbitrarily small probability of error from the observations gathered by the destination nodes at the end of Phase 3. The lemma is proven in Appendix 4.A.

Lemma 4.4.4. *The mutual information achieved by the $M \times M$ MIMO transmission between any two clusters is larger than $K_3 M$ for a constant $K_3 > 0$ and independent of M.*

The following lemma states that there is actually a way to encode the observations using fixed number of bits per observation and at the same time, not to degrade the performance of the overall channel significantly, that is, to still get a linear capacity growth for the resulting *quantized MIMO* channel. The proof of the lemma is given in Appendix 4.B.

Lemma 4.4.5. *There exists a strategy to encode the observations at a fixed rate Q bits per observation and get a linear growth κM of the mutual information for the resultant $M \times M$ quantized MIMO channel, for a constant $\kappa > 0$ independent of M and n.*

Putting it together, we have seen that the three phases described effectively realize virtual MIMO channels achieving spatial multiplexing gain M between the source and destination nodes in the network. Using these virtual MIMO

channels, each source is able to transmit ML bits in

$$\begin{aligned} t_{\text{total}} &= t_{\text{phase-1}} + t_{\text{phase-2}} + t_{\text{phase-3}} \\ &= \frac{18L}{K_1} M^{2-b} + 2Cn + \frac{18CQ}{K_1} M^{2-b} \end{aligned}$$

total channel uses where $L/C \geq \kappa$ for some $\kappa > 0$ independent of M (or n). This gives an aggregate throughput of

$$\begin{aligned} T &= \frac{nML}{(18L/K_1)M^{2-b} + 2Cn + (18CQ/K_1)M^{2-b}} \\ &\geq \frac{nM}{(18/K_1)M^{2-b} + (2/\kappa)n + (18Q/K_1\kappa)M^{2-b}} \\ &\geq K_2 n^{\frac{1}{2-b}} \end{aligned} \tag{4.9}$$

for some $K_2 > 0$ independent of n, by choosing $M = n^{\frac{1}{2-b}}$ with $0 \leq b < 1$, which is the optimal choice for the cluster size as a function of b. A failure arises if there are not order M nodes in each cluster or the scheme used in Phases 1 and 3 fails to achieve the promised throughput. Combining the result of Lemma 2.3.1-(e) with the computed failure probabilities for Phases 1 and 3 yields

$$P_f \leq 4ne^{-M^{c_1}} + \frac{n}{M} e^{-\Lambda(\delta)M} \leq e^{-n^{c_2}}$$

for some $c_2 > 0$.

In order to conclude the proof of Lemma 4.4.1, we should note that the new scheme achieves the same aggregate throughput scaling when the network experiences interference from the exterior. In phases 1 and 3, this external interference with $\text{INR}_{ex} \leq K_I \text{SNR}_l$ will simply add to the inter-cluster interference experienced by the nodes. Recall that the inter-cluster interference to noise power ratio is bounded by $\text{INR}_{ic} \leq K_{I_1} \text{SNR}(A_c)$ and since $\text{SNR}(A_c) > \text{SNR}_l$, the external interference is weaker than the inter-cluster interference. For the MIMO phase, the external interference will lead to uncorrelated background-noise-plus-interference at the receiving nodes which is not necessarily Gaussian. In Appendices 4.A and 4.B we prove the results stated in Lemma 4.4.4 and Lemma 4.4.5 for this more general case. This concludes the proof of Lemma 4.4.1. □

Proof of Lemma 4.4.2: The scheme that proves Lemma 4.4.2 is completely similar to the one described above. Lemma 4.4.3 states that when $\alpha = 2$, the inter-cluster interference to noise power ratio experienced during Phases 1 and 3 is upperbounded by $K_{I_2} \text{SNR}(A_c) \log n \leq K'_{I_2} \text{SNR}(A_c) \log M$. From the assumptions in the lemma, there is furthermore the external interference with INR bounded by $K_I \text{SNR}_l \log n$ which is adding to the inter-cluster interference. Under these conditions, the scheme in the hypothesis of Lemma 4.4.2 achieves an aggregate rate $K_1 \frac{M^{b_1}}{(\log M)^{b_2}}$ when used to handle the traffic in these phases.

For the second phase, we have the following lemma which provides a lower bound on the spatial multiplexing gain of the quantized MIMO channel under the external interference experienced. The proof is relegated to the end of Appendix 4.B.

Lemma 4.4.6. *Let the MIMO signal received by the nodes in the destination cluster be corrupted by an external interference of* $INR \leq K_I SNR_l \log M$*, uncorrelated over different nodes and independent of the transmitted signals. There exists a strategy to encode these corrupted observations at a fixed rate* Q *bits per observation and get a* $\kappa' M/\log M$ *growth of the mutual information for the resulting* $M \times M$ *quantized MIMO channel.*

A capacity of $\kappa' M/\log M$ for the resulting MIMO channel implies that there exists a code $\mathcal{C}$ that encodes L bits-long sub-blocks into $C \log M$ symbols, where $L/C \geq \kappa'$ for a constant $\kappa' > 0$, so that the transmitted bits can be decoded at the destination nodes with arbitrarily small probability of error, for L and C sufficiently large. Hence, starting again with a block of LM bits in each source node, the LM^2 bits in the first phase can be distributed in $(L/K_1)M^{2-b_1}(\log M)^{b_2}$ channel uses. In the second phase, the L bits-long sub-blocks now need to be encoded into $C \log M$ symbols, hence the transmission for each source-destination pair takes $C \log M$ channel uses. The whole phase takes $Cn \log M$ channel uses. In the third phase, there are now a total of $CM^2 \log M$ MIMO observations encoded into $CQM^2 \log M$ bits, that need to be exchanged between the nodes in the cluster. With the scheme of aggregate rate $K_1 \frac{M^{b_1}}{(\log M)^{b_2}}$, we need $(CQ/K_1)M^{2-b_1}(\log M)^{b_2+1}$ channel uses to complete the phase. Choosing $M = n^{\frac{1}{2-b_1}}$ gives an aggregate throughput of $K_2 n^{\frac{1}{2-b_1}}/(\log n)^{b_2+1}$ for the new scheme for a constant $K_2 > 0$ and independent of n. This concludes the proof of Lemma 4.4.2. □

We continue with the proof of the Lemma 4.4.3. Lemmas 4.4.4, 4.4.5 and 4.4.6 are proven in Appendices 4.A and 4.B.

Proof of Lemma 4.4.3: Consider a node v in cluster V operating under the 9-TDMA scheme in Figure 4.7. The interfering signal received by this node from the simultaneously operating clusters $\mathcal{U}_V$ is given by

$$I_v = \sum_{U \in \mathcal{U}_V} \sum_{k \in U} H_{vk} X_k$$

where H_{vk} are the channel coefficients given by (2.1) and X_k is the signal transmitted by node k which is located in a simultaneously operating cluster U. First note that the signals I_v and $I_{v'}$ received by two different nodes v and v' in V are uncorrelated since the channel coefficients H_{vk} and $H_{v'k}$ are

independent for all k. The power of the interfering signal I_v is given by

$$P_I = \sum_{U \in \mathcal{U}_V} \sum_{k \in U} \frac{GP_k}{(r_{vk})^\alpha}$$

where we used the fact that channel coefficients corresponding to different nodes k are independent. As illustrated by Figure 5, the interfering clusters $\mathcal{U}_V$ can be grouped based on their distance to V such that each group $\mathcal{U}_V(i)$ contains $8i$ clusters or less. All the clusters in group $\mathcal{U}_V(i)$ are separated by a distance larger than $(3i-1)\sqrt{A_c}$ from V for $i = 1, 2, \ldots$. Recall that A_c is the cluster area. The number of such groups can be simply bounded by the number of clusters n/M in the network. Thus,

$$\begin{aligned} P_I &< \sum_{i=1}^{n/M} \sum_{U \in \mathcal{U}_V(i)} \sum_{k \in U} \frac{GP_k}{((3i-1)\sqrt{A_c})^\alpha} \\ &\leq M \frac{GP}{(\sqrt{A_c})^\alpha} \sum_{i=1}^{n/M} 8i \frac{1}{(3i-1)^\alpha} \end{aligned} \tag{4.10}$$

where we have used the fact that the powers of the signals are bounded by the average power constraint P. The sum in (4.10) is convergent for $\alpha > 2$. This leads to

$$\mathrm{INR}_{ic} = \frac{P_I}{N_0 W} \leq K_{I_1} M \frac{GP}{N_0 W (\sqrt{A_c})^\alpha} = K_{I_1} \mathrm{SNR}(A_c).$$

For $\alpha = 2$, the sum can be bounded by $K_{I_2} \log n \, \mathrm{SNR}(A_c)$ where K_{I_2} is a constant independent of n. □

4.5 Power-Limited Hierarchical Cooperation

In the previous section, we have seen that when $\mathrm{SNR}_l \geq 1$, i.e., when the long-distance SNR in the network does not decrease to zero with increasing number of users in the network, distributed MIMO communication with hierarchical cooperation can achieve linear throughput scaling. In this section, we would like to evaluate the performance of the scheme in a power-limited setting when the long-distance SNR in the network vanishes with increasing n. If the scheme described in the previous section is used without modification in such power-limited networks, it leads to very poor performance. Observe from (4.23) in Appendix 4.A that the capacity of the long-range MIMO transmissions in the scheme scale like $M \, \mathrm{SNR}_l$ when $\mathrm{SNR}_l < 0$ dB. In other words, the per-node or per-stream MIMO capacity decreases like $\Theta(\mathrm{SNR}_l)$ with increasing system size n. Therefore the bits in each stream have to be encoded into longer and longer sequences of channel symbols. This is similarly the case for point-to-point MIMO communication in the power-limited regime. However, here the

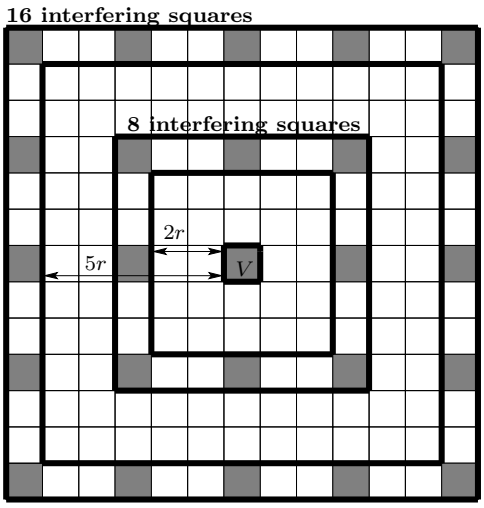

Figure 4.7: Grouping of interfering clusters in the 9-TDMA Scheme.

distributed nature of the receive cluster further degrades performance. The power-limited MIMO transmissions lead to the accumulation of many MIMO observations in the third phase, each one of these observations containing mostly noise with a signal component of very low power. These large number of observations have to be quantized and delivered to their destinations. As opposed to the $\mathrm{SNR}_l \geq 1$ case, the total traffic generated in the third phase becomes much larger than the traffic in the first phase. The performance of the overall scheme is limited by the burden of transporting a large number of observations that contain mostly noise in the third phase of the distributed MIMO scheme.

In this section, we propose two simple modifications of the hierarchical cooperation scheme for power-limited networks. Both of these modifications achieve a total throughput scaling of $n^{1-\epsilon}\mathrm{SNR}_l$ for any $\epsilon > 0$, when $\mathrm{SNR}_l < 0$ dB. Alternatively, the aggregate throughput can be expressed in terms of the nearest neighbor SNR in the network, $\mathrm{SNR}_s = n^{\beta}$, recalling the relation

$$\mathrm{SNR}_l = n^{1-\alpha/2}\mathrm{SNR}_s,$$

given earlier in (2.7). The aggregate throughput achieved by the modified hierarchical cooperation schemes in this section is lowerbounded by

$$K_\epsilon\, n^{2-\alpha/2+\beta-\epsilon}, \qquad \text{if} \qquad \beta < \alpha/2 - 1$$

for a constant $K_\epsilon > 0$ independent of n. The key to both modifications is to ensure that the MIMO observations in the third phase contain a constant amount of signal power that does not decrease to zero with increasing system size. Therefore, the burden of transporting a large number of excessively noisy observations in the third phase is removed.

The first modification we propose is to run the hierarchical cooperation scheme described in Section 4.4 as it is, a fraction SNR_l of the time with power P/SNR_l per node and remain silent for the rest of the time. (Recall that $\mathrm{SNR}_l < 1$.) This meets the given average power constraint P per node and achieves an aggregate throughput of

$$T \geq \mathrm{SNR}_l \times K_\epsilon \, n^{1-\epsilon},$$

since during operation the effective long-distance SNR in the network is larger than 0 dB. Therefore, the conditions of Theorem 4.2.1 are satisfied and aggregate throughput of $K_\epsilon \, n^{1-\epsilon}$ is achieved.

This transmission strategy creates a "bursty" hierarchical cooperation scheme with a high peak-to-average power ratio. (The idea of using burstiness in improving the low-SNR performance of relay networks was introduced in [23] in the context of single-relay networks.) However, although we talk in terms of time in the above discussion, such burstiness can just as well be implemented over frequency with only a fraction of the total bandwidth W used. For example, this can be implemented in an OFDM system, using a subset of the sub-carriers at any one time, but putting more energy in the active sub-carriers. This way, the peak power remains constant over time.

The alternative to bursty communication is repetition coding. However, this second strategy can only be applied if the coherence time of the channel fading process is long enough. Consider the channel model given in Section 2, where the signal received by node i at time m is given by

$$Y_i[m] = \sum_{k \neq i} H_{ik}[m] X_k[m] + Z_i[m] \tag{4.11}$$

where Z_i is *white* Gaussian noise. Let us assume the channel coefficients $H_{ik}[m]$ remain constant for a duration of $N = 1/\mathrm{SNR}_l$ channel uses, i.e,

$$H_{ik}[m] = H_{ik}, \qquad \text{for} m = 1, \ldots, N.$$

Let each transmission in the network be exactly repeated over these $N = 1/\mathrm{SNR}_l$ channel uses. Each receiving node i can combine the signals it observes during these N channel uses and obtain the signal

$$\begin{aligned} \tilde{Y}_i[1] &= \frac{1}{N} \sum_{l=1}^{N} \sum_{k \neq i} H_{ik} X_k[l] + \frac{1}{N} \sum_{l=1}^{N} Z_i[l] \\ &= \sum_{k \neq i} H_{ik} X_k[1] + \frac{1}{N} \sum_{l=1}^{N} Z_i[l]. \end{aligned} \tag{4.12}$$

Since the noise is white, the power of the noise in $\tilde{Y}_i$ is reduced by a factor of $1/N$ as compared to Y_i in (4.11). Equivalently, the SNR is increased by a

factor of N. For the network described by the channel model in (4.12), the long distance SNR is given by $N \times \mathrm{SNR}_l = 1$, therefore Theorem 4.2.1 applies and an aggregate throughput of $K_\epsilon\, n^{1-\epsilon}$ is achieved. Because of the repetitions, this yields an aggregate throughput

$$\frac{1}{N} \times K_\epsilon n^{1-\epsilon} = K_\epsilon n^{1-\epsilon} \mathrm{SNR}_l$$

for the original network. Note that the quantity $n\mathrm{SNR}_l$ can be interpreted as the total power transferred between a size n transmit cluster and a size n receive cluster separated by a distance of the order of the diameter of the network. This power transfer is taking place at the top level of the hierarchy (see Figure 1.1). The fact that the achievable rate is proportional to the power transfer further emphasizes that the scheme is power-limited when $\mathrm{SNR}_l < 0$ dB as opposed to degrees-of-freedom limited in the previous case of $\mathrm{SNR}_l > 0$ dB.

4.6 The MIMO-Multihopping Scheme

In this section, we combine the hierarchical cooperation scheme presented in Section 4.4 with the multi-hopping scheme in Section 4.3. We divide the network into square cells of area A_c and relay the packets between the source-destination pairs by hopping from one cell to the next, while each hop is performed by distributed MIMO communication and hierarchical cooperation. Let $M = A_c n/A$ be the average number of nodes contained in each cell. Later we will argue more precisely that for our particular choice of A_c, Lemma 2.3.1-(e) ensures that there are $\Theta(M)$ nodes in all cells w.h.p. As in the case of pure multi-hopping, we follow a simplistic route between the source-destination pairs by first relaying the packets horizontally and then vertically as shown in Figure 4.8. Hence, the relaying burden imposed on a given cell is due to the source nodes that lie in its horizontal slab and destination nodes that lie in its vertical slab. The number of nodes contained in a slab of area $\sqrt{A_c\, A}$ is $\Theta(\sqrt{Mn})$. Hence, there can be at most $O(2\sqrt{Mn})$ source-destination routes that pass from a given cell. Let us randomly associate each of the source-destination pairs whose routes pass through a given cell with one of the M nodes in this cell so that each node is associated with at most $O(2\sqrt{n/M})$ source-destination pairs. The only rule that we need to respect while doing this association is that if a source-destination route starts or ends in a certain cell, then the node associated to this source-destination pair should naturally be its respective source or destination node. The nodes associated to a source-destination pair are those that will decode, temporarily store and forward the packets of this source-destination pair during the multi-hop operation. The following lemma states a key result regarding the rate of transmission between neighboring cells.

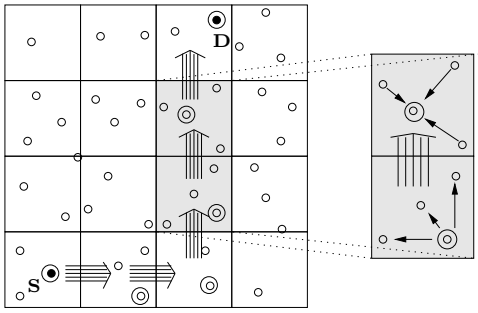

Figure 4.8: The figure illustrates the optimal scheme in Regime IV which is based on cooperating locally and multi-hopping globally. Note that packets are transmitted by multi-hopping on the network level and each hop is realized with distributed MIMO transmissions combined with hierarchical cooperation.

Lemma 4.6.1. *If $SNR(A_c) \geq 1$, there exists a strategy (based on distributed MIMO communication with hierarchical cooperation) that allows each node in the network to relay its packets to their respective destination nodes in the adjacent cells at a rate*

$$R_{relay} \geq K_\epsilon \, M^{-\epsilon}$$

for any $\epsilon > 0$ and a constant $K_\epsilon > 0$ independent of M.

In steady-state operation, the outbound rate of a relay node given in the lemma should be shared between the $O(2\sqrt{n/M})$ source-destination pairs that the relay is responsible for. Hence, the rate per source-destination pair in the network is lowerbounded by

$$R \geq \frac{K_\epsilon}{2} \frac{M^{1/2-\epsilon}}{\sqrt{n}} \tag{4.13}$$

or equivalently, the aggregate rate achieved by the scheme is

$$T_{multihopping+HC} \geq \frac{K_\epsilon}{2} \sqrt{n} \, M^{1/2-\epsilon}.$$

Note that combining multi-hopping with hierarchical cooperation provides a-$\sqrt{M}$-fold-gain in the aggregate throughput as compared to pure multi-hopping, which indeed corresponds to $M = 1$ in the above discussion. Choosing larger M yields a larger aggregate throughput since it reduces the relaying burden. Indeed if we could choose $M = n$, we could get linear scaling in which case the scheme would reduce to pure hierarchical cooperation. However since $\mathrm{SNR}_l < 1$, the condition $\mathrm{SNR}(A_c) \geq 1$ is not satisfied for $A_c = A$, so M can not be as large as n. The largest cluster area that satisfies the condition $\mathrm{SNR}(A_c) \geq 1$ is,

$$A_c = \frac{A}{n} \mathrm{SNR}_s^{1/(\alpha/2-1)}. \tag{4.14}$$

Note that this choice for A_c coincides with the intuition suggested by the upper bound derivation, discussed in Section 3.4. For $\mathrm{SNR}_s > 1$ and $\mathrm{SNR}_l < 1$, the choice for A_c yields $A/n < A_c < A$ as expected. If we assume $\mathrm{SNR}_s = n^\beta$ with $0 < \beta < \alpha/2-1$, each cell contains $\Theta(M) = \Theta(\mathrm{SNR}_s^{1/(\alpha/2-1)})$ nodes w.h.p. by Lemma 2.3.1-(e). (When SNR_s is a constant larger than 1, or $\mathrm{SNR}_s = \log n$, part(e) of Lemma 2.3.1 does not ensure that the number of nodes in each cell concentrate around their mean value for large n. This technical difficulty can be overcome by allowing a cluster area $A_c n^{\epsilon_1}$ for arbitrarily small $\epsilon_1 > 0$. This choice for the cluster area will ensure that there are $\Theta(M)$ nodes in all cells and will degrade the aggregate throughput scaling of the scheme only by $n^{-\epsilon_1}$. Thus the following result can indeed be established for any $\mathrm{SNR}_s > 1$ and $\mathrm{SNR}_l < 1$.)

With the choice (4.14) for A_c, we have

$$T_{multihop+HC} \geq K_{\epsilon'} \sqrt{n}\, \mathrm{SNR}_s^{\frac{1}{\alpha-2}-\epsilon'} \quad \text{if } \mathrm{SNR}_s > 1 \text{ and } \mathrm{SNR}_l < 1 \tag{4.15}$$

for any $\epsilon' > 0$. In terms of scaling exponents, we have

$$e_{multihop+HC}(\alpha, \beta) = 1/2 + \beta/(\alpha - 2) \quad \text{if } 0 < \beta < \alpha/2 - 1$$

which matches the upper bound (3.16) in the third regime.

Proof of Lemma 4.6.1: Let us concentrate only on two neighboring cells in the network. (Consider for example the two cells highlighted in Fig. 4.8): The two neighboring cells together form a network of $2M$ nodes randomly and uniformly distributed on a rectangular area $2\sqrt{A_c} \times \sqrt{A_c}$. Let the M nodes in one of the cells be sources and the M nodes in the other cell be destinations and let these source and destination nodes be paired up arbitrarily to form M S-D pairs. (This traffic will later be used to model the hop between two adjacent cells.) If $\mathrm{SNR}(A_c) > 1$, Theorem 4.2.1 ensures that with hierarchical cooperation an aggregate rate

$$M\, R_{relay} \geq K_\epsilon\, M^{1-\epsilon} \tag{4.16}$$

is achievable for these M source destination pairs. Note that by Remark 4.4.1, this rate is also achievable under external interference which is zero mean and uncorrelated across nodes and with $\mathrm{INR}_{ex} \leq K_I \mathrm{SNR}(A_c)$ for a constant K_I.

Now, let us turn back to our original problem concerning the steady-state operation of the multi-hop scheme. At each hop, each of the M nodes in a cell need to relay its packets to one of the four (left, right, up and down) adjacent cells. Since the source-destination routes are randomly assigned to the nodes in the cell, there are $M/4$ nodes on the average that want to transmit in each direction. These transmissions can be realized successively using hierarchical cooperation and the relaying rate in (4.16) can be achieved in each transmission. On the other hand, the TDMA between the four transmissions will reduce the overall relaying rate by a factor of 4. Moreover, one should also consider a

TDMA scheme between the cells allowing only those cells that are sufficiently separated in space to operate simultaneously so that the inter-cell-interference in the network does not degrade the quality of the transmissions significantly. Employing a 9-TDMA scheme as we did in the case of multi-hopping, the inter-cell interference can be shown to satisfy $\mathrm{INR}_{ic} < K_{I_1}\mathrm{SNR}(A_c)$ by a simple modification of Lemma 4.4.3. The 9-TDMA scheme will further reduce the rate by a factor of 9 in (4.16) however will not affect the scaling law in (4.15). □

4.A Linear Scaling Law for the MIMO Channel

The discrete-time baseband equivalent $M \times M$ MIMO channel between two clusters S and D is given by

$$Y = HX + I + Z, \tag{4.17}$$

where Y is an $M \times 1$ random vector representing the signals received by nodes D, X is the $M \times 1$ vector of transmitted signals from S and the entries H_{ik} of the $M \times M$ channel matrix H are given in (2.1). Recall that Z represents the additive white Gaussian noise of variance N_0 and I is $M \times 1$ vector with uncorrelated entries of variance P_I/W, representing the external interference signals experienced by nodes in D and satisfying

$$\mathrm{INR}_{ex} = \frac{P_I}{N_0 W} \leq K_I \,\mathrm{SNR}_l \tag{4.18}$$

Assume that the transmitted signals X_i are independent random variables with distribution $\sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ of variance

$$\sigma^2 = \frac{nP}{MW}.$$

Recall our discussion in Section 4.4.1 that the nodes in S are transmitting with power nP/M during the MIMO transmissions. It is well known that the achievable mutual information is lower bounded by assuming that the interference-plus-noise $I + Z$ vector is independent across nodes in D and complex circularly-symmetric Gaussian distributed. (See for example Theorem 5 of [15] for a precise statement and proof of this in the MIMO case.)

Let the distance between the midpoints of the two clusters S and D be r_{SD}. With our transmission strategy in the MIMO phase, there exists $b > a > 0$ with a and b independent of n, such that $r_{ik}^{-\alpha/2} = r_{SD}^{-\alpha/2}\,\rho_{ik}$, where all ρ_{ik} lie in the interval $[a, b]$. This is true both in the cases when S and D are neighboring clusters and when they are not. In the case when S and D are not neighboring clusters, observe that for any $i \in D$, $k \in S$

$$r_{SD} - \sqrt{2A_c} \leq r_{ik} \leq r_{SD} + \sqrt{2A_c},$$

while $r_{SD} \geq 2\sqrt{A_c}$. These two relations yield

$$\left(\frac{\sqrt{2}}{\sqrt{2}+1}\right)^{\alpha/2} \leq \left(\frac{r_{SD}}{r_{ik}}\right)^{\alpha/2} \leq \left(\frac{\sqrt{2}}{\sqrt{2}-1}\right)^{\alpha/2}. \tag{4.19}$$

When S and D are neighboring clusters, $r_{SD} \leq r_{ik} \leq r_{SD} + \sqrt{2A_c}$ and $r_{SD} \geq \sqrt{A_c}$ which leads to

$$\left(\frac{\sqrt{1}}{\sqrt{2}+1}\right)^{\alpha/2} \leq \left(\frac{r_{SD}}{r_{ik}}\right)^{\alpha/2} \leq 1.$$

We can rewrite the relation (4.17) as

$$Y \;=\; (\sqrt{G}\, r_{SD}^{-\alpha/2})\; FX \;+\; I \;+\; Z, \tag{4.20}$$

where $F_{ik} = \rho_{ik} \exp(j\,\theta_{ik})$. The capacity of this MIMO channel is lower bounded by the capacity of the channel

$$Y' \;=\; \sqrt{G}\,(\sqrt{A})^{-\alpha/2}\; FX \;+\; I \;+\; Z, \tag{4.21}$$

since the separation between any two clusters is upper bounded by the diameter of the network, $r_{SD} \leq \sqrt{A}$. By assuming perfect channel state information at the receiver side, the mutual information of the above MIMO channel is bounded from below by

$$\begin{aligned} I(X;Y',H) &\geq \mathbb{E}\left(\log\det\left(I + G(\sqrt{A})^{-\alpha}\frac{nP/M}{N_0W+P_I}\,FF^*\right)\right) \\ &= \mathbb{E}\left(\log\det\left(I + \frac{\mathrm{SNR}_l}{1+\mathrm{INR}_{ex}}\,\frac{1}{M}\,FF^*\right)\right) \end{aligned}$$

since the separation between the two clusters is upper bounded by the diameter of the network, $r_{SD} \leq \sqrt{A}$. We furthermore obtain

$$I(X;Y',H) \geq \mathbb{E}\left(\log\det\left(I + \frac{\mathrm{SNR}_l}{1+K_I\mathrm{SNR}_l}\,\frac{1}{M}\,FF^*\right)\right), \tag{4.22}$$

using the upperbound (4.18).

Let λ be chosen uniformly among the M eigenvalues of $\frac{1}{M}FF^*$. The above lowerbound on the mutual information can be written as

$$\begin{aligned} I(X;Y',H) &\geq M\,\mathbb{E}\left(\log\left(1 + \frac{\mathrm{SNR}_l}{1+K_I\mathrm{SNR}_l}\,\lambda\right)\right) \\ &\geq M\,\log\left(1 + \frac{\mathrm{SNR}_l}{1+K_I\mathrm{SNR}_l}\,t\right)\,\mathbb{P}(\lambda > t) \end{aligned}$$

for any $t \geq 0$. By the Paley-Zygmund inequality, if $0 \leq t < \mathbb{E}(\lambda)$, we have

$$\mathbb{P}(\lambda > t) \geq \frac{(\mathbb{E}(\lambda) - t)^2}{\mathbb{E}(\lambda^2)}.$$

We therefore need to compute both $\mathbb{E}(\lambda)$ and $\mathbb{E}(\lambda^2)$. We have,

$$\begin{aligned}\mathbb{E}(\lambda) &= \frac{1}{M}\,\mathbb{E}\left(\mathrm{Tr}\left(\frac{1}{M}\,FF^*\right)\right)\\ &= \frac{1}{M^2}\sum_{i,k=1}^{M}\mathbb{E}(|F_{ik}|^2)\\ &= \frac{1}{M^2}\sum_{i,k=1}^{M}\rho_{ik}^2 \geq a^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\lambda^2) &= \frac{1}{M}\mathbb{E}\left(\mathrm{Tr}\left(\frac{1}{M^2}FF^*FF^*\right)\right)\\ &= \frac{1}{M^3}\sum_{iklm=1}^{M}\mathbb{E}(F_{ik}\overline{F_{lk}}F_{lm}\overline{F_{im}})\\ &\leq \frac{2}{M^3}\sum_{ikm=1}^{M}\mathbb{E}(|F_{ik}|^2)\,\mathbb{E}(|F_{im}|^2)\\ &= \frac{2}{M^3}\sum_{ikm=1}^{M}\rho_{ik}^2\,\rho_{im}^2 \leq 2b^4\end{aligned}$$

so $\mathbb{E}(\lambda) \geq a^2$ and $\mathbb{E}(\lambda^2) \leq 2b^4$. This leads us to the conclusion that for any $t < a$, we have

$$I(X;Y',H) \geq M\,\log\left(1+\frac{\mathrm{SNR}_l}{1+K_I\mathrm{SNR}_l}\,t\right)\,\frac{(a^2-t)^2}{2b^4}, \tag{4.23}$$

Choosing e.g. $t = a/2$ and recalling that by Lemma 4.4.1 we have $\mathrm{SNR}_l \geq K_S$, shows that $I(X;Y,H)$ grows at least linearly with M. □

Lemma 4.A.1. *(Paley-Zygmund Inequality) Let X be a non-negative random variable such that $\mathbb{E}(X^2) < \infty$. Then for any $t \geq 0$ such that $t < \mathbb{E}(X)$, we have*

$$\mathbb{P}(X > t) \geq \frac{(\mathbb{E}(X)-t)^2}{\mathbb{E}(X^2)}$$

Proof: By the Cauchy-Schwarz inequality, we have for any $t \geq 0$:

$$\mathbb{E}(X\,1_{X>t}) \leq \sqrt{\mathbb{E}(X^2)\,\mathbb{P}(X>t)}$$

and also, if $t < \mathbb{E}(X)$,

$$\mathbb{E}(X\,1_{X>t}) = \mathbb{E}(X) - \mathbb{E}(X\,1_{X\leq t}) \geq \mathbb{E}(X) - t > 0$$

Therefore,

$$\mathbb{P}(X > t) \geq \frac{(\mathbb{E}(X) - t)^2}{\mathbb{E}(X^2)}.$$

□

Note that the achievability results in this paper can be extended to the slow fading case, provided that Lemma 4.4.4 can be proved in the slow fading setting. In that case, one would need to show that the expression inside the expectation in (4.22) concentrates around its mean exponentially fast in M. However, another difficulty might arise from the lack of averaging of the phases in the interference term, which leads to a non-spatially decorrelated noise term Z. Although proving the result might require some technical effort, we believe it holds true, due to the self-averaging effect of a large number of independent random variables.

4.B Achievable Rates on Quantized Channels

In order to conclude the discussion on the throughput achieved by our scheme, we need to show that the quantized MIMO channel achieves the same spatial multiplexing gain as the MIMO channel. In Theorem 4.B.1 below, we give a simple achievability region for general quantized channels. (Note that a stronger result is established in [31, Theorem 3] that implies Theorem 4.B.1 as a special case.) The required result for the quantized MIMO channel is then found as an easy application of Theorem 4.B.1. We start by formally defining the general quantized channel problem in a form that is of interest to us and proceed with several definitions that will be needed in the sequel.

Let us consider a discrete-time memoryless channel with single input of alphabet $\mathcal{X}$ and M outputs of respective alphabets $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$. The channel is statistically described by a conditional probability distribution $p(y_1, \ldots, y_M|x)$ for each $y_1 \in \mathcal{Y}_1, \ldots, y_M \in \mathcal{Y}_M$ and $x \in \mathcal{X}$. The outputs of the channel are to be followed by quantizers which independently map the output alphabets $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$ to the respective reproduction alphabets $\hat{\mathcal{Y}}_1, \ldots, \hat{\mathcal{Y}}_M$. The aim is to recover the transmitted information through the channel by observing the outputs of the quantizers. Communication over the channel takes place in the following manner: a message W, drawn from the index set $\{1, 2, \ldots, L\}$ is encoded into a codeword $X^m(W) \in \mathcal{X}^m$, which is received as M random sequences $(Y_1^m, \ldots, Y_M^m) \sim p(y_1^m, \ldots, y_M^m|x^m)$ at the outputs of the channel. The quantizers themselves consist of encoders and decoders, where the i'th encoder describes its corresponding received sequence Y_i^m by an index $U_i(Y_i^m) \in \{1, 2, \ldots, L_i\}$, and decoder i represents Y_i^m by an estimate $\hat{Y}_i^m(U_i) \in \hat{\mathcal{Y}}_i^m$. The channel decoder then observes the reconstructed sequences $\hat{Y}_1^m, \ldots, \hat{Y}_M^m$ and guesses the index W by an appropriate decoding rule $\hat{W} = g(\hat{Y}_1^m, \ldots, \hat{Y}_M^m)$. An error occurs if $\hat{W}$ is not the same as the index W that was transmitted. The complete model under investigation is shown in Fig. 8. An $(L; L_1, \ldots, L_M; m)$ code for this channel is a joint (L, m)

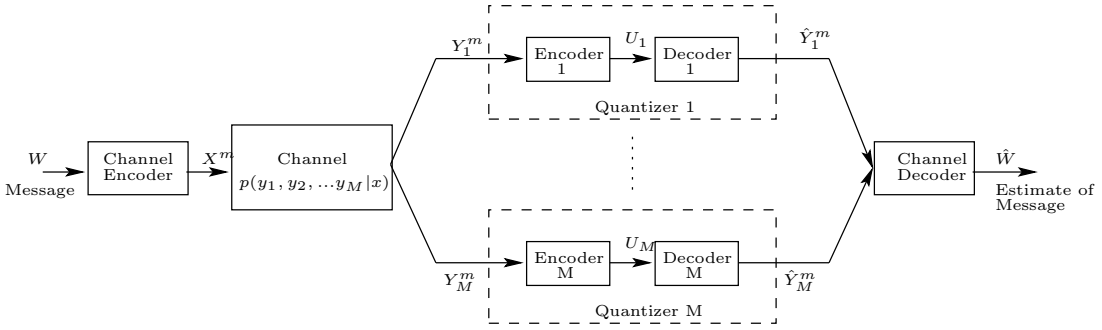

Figure 4.9: The Quantized Channel Problem.

channel code and M quantization codes $(L_1, m), \ldots, (L_M, m)$; more specifically, it is two sets of encoding and decoding functions, the first set being the channel encoding function $X^m : \{1, 2, \ldots, L\} \to \mathcal{X}^m$ and the channel decoding function $g : \hat{\mathcal{Y}}_1^m \times \cdots \times \hat{\mathcal{Y}}_M^m \to \{1, 2, \ldots, L\}$, and the second set consists of the encoding functions $U_i : \mathcal{Y}_i^m \to \{1, 2, \ldots, L_i\}$ and decoding functions $\hat{Y}_i^m : \{1, 2, \ldots, L_i\} \to \hat{\mathcal{Y}}_i^m$ for $i = 1, \ldots, M$ used for the quantizations. We define the (average) probability of error for the $(L; L_1, \ldots, L_M; m)$ code by

$$P_e^m = \frac{1}{L} \sum_{w=1}^{L} \mathbb{P}(\hat{W} \neq w \mid W = w).$$

A set of rates $(R; R_1, \ldots, R_M)$ is said to be achievable if there exists a sequence of
$(2^{mR}; 2^{mR_1}, \ldots, 2^{mR_M}; m)$ codes with $P_e^m \to 0$ as $m \to \infty$. Note that determining achievable rates $(R; R_1, \ldots, R_M)$ is not a trivial problem, since there is trade-off between maximizing R and minimizing $R_1, \ldots, R_M$.

Theorem 4.B.1. *(Achievability for the Quantized Channel Problem) Given a probability distribution $q(x)$ on $\mathcal{X}$ and M conditional probability distributions $q_j(\hat{y}_j|y_j)$ where $y_j \in \mathcal{Y}_j$, $\hat{y}_j \in \hat{\mathcal{Y}}_j$ and $j = 1, \ldots, M$, all rates $(R; R_1, \ldots, R_M)$ such that $R < I(X; \hat{Y}_1, \ldots, \hat{Y}_M)$ and $R_j > I(Y_j; \hat{Y}_j)$ are achievable. Specifically, given any $\delta > 0$, $q(x)$ and $q_j(\hat{y}_j|y_j)$, together with rates $R < I(X; \hat{Y}_1, \ldots, \hat{Y}_M)$ and $R_j > I(Y_j; \hat{Y}_j)$ for $j = 1, \ldots, M$; there exists a $(2^{mR}; 2^{mR_1}, \ldots, 2^{mR_M}; m)$ code such that $P_e^m < \delta$.*

Proof: The proof of the theorem for discrete finite-size alphabets relies on a random coding argument based on the idea of joint (*strong*) typicality. For the idea of strong typicality and properties of typical sequences, see [12, Ch. 13.6]. The proof can be outlined as follows. Given $q(x)$ generate a random channel codebook $\mathcal{C}_c$ with 2^{mR} codewords, each of length m, independently from the

distribution

$$q(x^m) = \prod_{k=1}^{m} q(x^m(k)).$$

and call them $X^m(1), X^m(2), \ldots, X^m(2^{mR})$. Also generate M quantization codebooks $\mathcal{C}_i, i = 1, \ldots, M$, each codebook $\mathcal{C}_i$ consisting of 2^{mR_i} codewords drawn independently from

$$p_i(\hat{y}_i^m) = \prod_{k=1}^{m} \sum_{\substack{x \in \mathcal{X} \\ y_1 \in \mathcal{Y}_1, \ldots, y_M \in \mathcal{Y}_M}} q(x) p(y_1, \ldots, y_M | x) q_i(\hat{y}_i^m(k) | y_i).$$

and index them as $\hat{Y}_i^m(1), \hat{Y}_i^m(2), \ldots \hat{Y}_i^m(2^{mR_i})$. Given the message w, send the codeword $X^m(w)$ through the channel. The channel will yield $Y_1^m, \ldots, Y_M^m$. Given the channel output Y_i^m at the i'th quantizer, choose j_i such that $(Y_i^m, \hat{Y}_i^m(j_i))$ are jointly typical. If there exists no such j_i, declare an error. If the number of codewords in the quantization codebook 2^{mR_i} is greater than $2^{mI(Y_i;\hat{Y}_i)}$, the probability of finding no such j_i decreases to zero exponentially as m increases. The probability of failing to find such an index in at least one of the M quantizers is bounded above by the union bound with the sum of M exponentially decreasing probabilities in m. Given $\hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M)$ at the channel decoder, choose the unique $\hat{w}$ such that $(X^m(\hat{w}), \hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M))$ are jointly typical. The fact that $(X^m(w), \hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M))$ will be jointly typical with high probability can be established by identifying the Markov chains in the problem and applying Markov Lemma [12, Lemma 14.8.1] repeatedly. Observing that $(Y_1^m, \ldots, Y_M^m, \hat{Y}_1^m, \ldots, \hat{Y}_i^m) - Y_{i+1}^m - \hat{Y}_{i+1}^m$ forms a Markov chain and recursively applying Markov Lemma, we conclude that $(Y_1^m, \ldots, Y_M^m, \hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M))$ are jointly typical with probability approaching 1 as m increases. Observing that $X^m - (Y_1^m, \ldots, Y_M^m) - (\hat{Y}_1^m, \ldots, \hat{Y}_M^m)$ forms another Markov chain, we have $(X^m(w), \hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M))$ jointly typical with high probability again by Markov Lemma. If there are more than one codewords X^m that are jointly typical with $(\hat{Y}_1^m(j_1), \ldots, \hat{Y}_M^m(j_M))$, we declare an error. The probability of having more than one such sequence will decrease exponentially to zero as m increases, if the number of channel codewords 2^{mR} is less than $2^{mI(X;\hat{Y}_1,\ldots,\hat{Y}_M)}$. Hence if $R < I(X; \hat{Y}_1, \ldots, \hat{Y}_M)$ and $R_i > I(Y_i; \hat{Y}_i)$, the probability of error averaged over all codes decreases to zero as $m \to \infty$. This shows the existence of a code that achieves rates $(R; R_1, \ldots, R_M)$ with arbitrarily small probability of error. The result can be readily extended to memoryless channels with discrete-time and continuous alphabets by standard arguments (see [20, Ch.7]). □

Proof of Lemma 4.4.5: Now we turn to our original problem. We need to show that it is possible to encode the observations at the outputs of the MIMO channel at a fixed rate, while preserving the spatial multiplexing gain of the MIMO channel. This fact follows easily from Theorem 4.B.1: From

(4.20), the received signal at node $i \in D$ is given by

$$Y_i \;=\; (\sqrt{G}\, r_{SD}^{-\alpha/2}) \sum_{i \in S} F_{ik} X_k \;+\; I_i \;+\; Z_i.$$

Let us first scale the received signal to obtain

$$\begin{aligned}
\tilde{Y}_i &= \left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} Y_i \\
&= \sqrt{G}\,(\sqrt{A})^{-\alpha/2} \sum_{k \in S} F_{ik} X_k \;+ \left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} I_i \;+\; \left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} Z_i \\
&= \sqrt{G}\,(\sqrt{A})^{-\alpha/2} \sum_{k \in S} F_{ik} X_k \;+ \tilde{I}_i \;+ \tilde{Z}_i.
\end{aligned}$$

Next, we define the conditional probability densities

$$q_i(\hat{y}_i|\tilde{y}_i) = \mathcal{N}_{\mathbb{C}}(\hat{y}_i, \Delta^2) \tag{4.24}$$

for the quantization process. From Theorem 4.B.1 we know that for any distribution $p(x)$ on the input space, all rate pairs $(R; R_1, \ldots, R_M)$ are simultaneously achievable if

$$R_i > I(\tilde{Y}_i; \hat{Y}_i),\; i = 1, \ldots, M \quad \text{and} \quad R < I(X; \hat{Y}_1, \ldots, \hat{Y}_M)$$

where now R_i is the encoding rate of the i'th stream and R is the total transmission rate over the MIMO channel. With the distribution in (4.24), we have

$$I(\tilde{Y}_i; \hat{Y}_i) \leq \log\left(1 + \frac{\mathbb{E}(|\tilde{Y}_i|^2)}{\Delta^2}\right) \tag{4.25}$$

for any probability distribution $p(x)$ on the input space, where $\mathbb{E}(|\tilde{Y}_i|^2)$ is the variance of the rescaled observation $\tilde{Y}_i$. Note that if this variance is increasing with n, we should allow the variance of the quantization error, Δ^2, to also increase with n, in order to be able to keep the quantization rate R_i constant. The variance of $\tilde{Y}_i$ is given by

$$\mathbb{E}\left(|\tilde{Y}_i|^2\right) = G(\sqrt{A})^{-\alpha} \sum_{k \in S} \rho_{ik}^2 \frac{nP}{MW} \;+\; \left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} \frac{P_I}{W} \;+\; \left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} N_0.$$

Using the upper bound in (4.19) for ρ_{ik} and noting that $\left(\frac{\sqrt{A}}{r_{SD}}\right)^{-\alpha/2} \leq 1$ yields,

$$\begin{aligned}
\frac{\mathbb{E}\left(|\tilde{Y}_i|^2\right)}{N_0} &\leq \mathrm{SNR}_l \left(\frac{\sqrt{2}}{\sqrt{2}-1}\right)^{\alpha} + \mathrm{INR}_{ex} + 1 \\
&\leq \mathrm{SNR}_l \left(\frac{\sqrt{2}}{\sqrt{2}-1}\right)^{\alpha} + K_I \mathrm{SNR}_l + 1.
\end{aligned} \tag{4.26}$$

Thus, choosing

$$\frac{\Delta^2}{N_0} \leq K_Q \mathrm{SNR}_l \tag{4.27}$$

for a constant $K_Q > 0$ independent of M and n yields a constant upper bound on (4.25),

$$I(\tilde{Y}_i; \hat{Y}_i) \leq \log \left(1 + \frac{\left(\frac{\sqrt{2}}{\sqrt{2}-1}\right)^{\alpha} + K_I + \frac{1}{K_S}}{K_Q} \right) := R_Q$$

where we also make use of the fact that $\mathrm{SNR}_l \geq K_S$. So if we choose

$$R_i = R_Q + \varepsilon \qquad \forall i = 1, \ldots, M$$

for some $\varepsilon > 0$, all rates

$$R \leq I(X; \hat{Y}_1, \ldots, \hat{Y}_M)$$

are achievable on the quantized MIMO channel for any input distribution $p(x)$. Note that now the channel from X to $\hat{Y}_1, \ldots, \hat{Y}_M$ is given by

$$\hat{Y} = \sqrt{G}\,(\sqrt{A})^{-\alpha/2}\, F\, X \, + \tilde{I} \, + \tilde{Z} \, + D$$

where $D \sim \mathcal{N}_{\mathbb{C}}(0, \Delta^2 I)$ with Δ^2 specified in (4.27) is the quantization noise which is independent of all the other variables. This channel is equivalent to the MIMO channel in (4.21) and linear scaling of the capacity is established in Section 4.A. □

Proof of Lemma 4.4.6: Consider the case where the MIMO signals are corrupted by external interference with $\mathrm{INR}_{ex} \leq K_I \mathrm{SNR}_l \log M$. We would like to show that it is possible to quantize the observations at a fixed rate independent of M, or n and get $M/\log M$ capacity scaling for the resultant quantized MIMO channel. The proof follows by a small modification of the proof of Lemma 4.4.5. Following the same approach till (4.26), yields

$$\frac{\mathbb{E}\left(|\tilde{Y}_i|^2\right)}{N_0} \leq \mathrm{SNR}_l \left(\frac{\sqrt{2}}{\sqrt{2}-1} \right)^{\alpha} + K_I \mathrm{SNR}_l \log M + 1.$$

To be able to encode the observations at constant rates R_i, we need to choose

$$\frac{\Delta^2}{N_0} \leq K_Q \mathrm{SNR}_l \log M. \tag{4.28}$$

The channel from X to $\hat{Y}_1, \ldots, \hat{Y}_M$ is then given by

$$\hat{Y} = \sqrt{G}\,(\sqrt{A})^{-\alpha/2}\, F\, X \, + \tilde{I} \, + \tilde{Z} \, + D$$

where both $\tilde{I}$ and D have larger power to noise ratios $K_I \mathrm{SNR}_l \log M$ and $K_Q \mathrm{SNR}_l \log M$ respectively, instead of $K_I \mathrm{SNR}_l$ and $K_Q \mathrm{SNR}_l$. This yields an extra $\log M$ factor in the denominator of (4.22) which in turn yields $M/\log M$ capacity scaling for the quantized MIMO channel. □

Throughput-Delay Trade-off for Hierarchical Cooperation

5

In the previous two chapters, we have seen how scaling law results can be used to identify operating regimes of large wireless networks and to devise scaling optimal communication schemes for each regime. In this chapter, we use a scaling law formulation to highlight the qualitative properties of specific designs. More precisely, we derive a throughput-delay trade-off for the hierarchical cooperation scheme presented in Section 4.4.

The analysis of the delay performance of the hierarchical cooperation scheme reveals a key drawback of the architecture in Section 4.4 from the bulk-size point of view. The bulk-size of a scheme is defined as the minimum number of bits that should be communicated between each source-destination pair under this scheme. A modification to the architecture is devised in this chapter to overcome this major drawback. The key ingredient is a more careful study of the cooperation problem.

5.1 Introduction and Literature Overview

In the previous chapters, we have extensively discussed the results of the seminal work [27] of Gupta and Kumar in 2000, that has initiated the scaling law approach to wireless networks. We have seen that one of the main results of [27] is to establish the aggregate throughput scaling of the multi-hopping scheme as $\Theta(\sqrt{n})$. This has been interpreted as a rather negative result since it implies that the rate per source destination pair should decrease to 0 as $1/\sqrt{n}$ when n is large. However, the same result can also be interpreted as good news when compared to the performance of simple TDMA. If nodes cooperate and relay packets by multi-hopping from one node to the next, an aggregate throughput scaling of $\Theta(\sqrt{n})$ is achieved, when the simple scheme of time-sharing between direct transmissions from source nodes to destinations achieves only an aggre-

gate throughput $\Theta(1)$. The price to pay, however, is in terms of delay. In the multi-hopping scheme, the packets need to be retransmitted many times before they reach their actual destinations, which results in larger end-to-end delay. More precisely, as shown later in [21, 22], in a multi-hop scheme, bits are delivered to their destinations in $\Theta(\sqrt{n})$ time-slots on average after they leave their source nodes, while the average delay for the simple TDMA scheme remains only $\Theta(1)$. Note that this accounts only for on-the-flight delay; the queuing delay at the source node is not considered.

In the previous chapter, we have introduced a hierarchical cooperation architecture for distributed MIMO communication that achieves an aggregate throughput scaling arbitrarily close to linear, i.e. $T_h(n) = \Theta(n^{\frac{h}{h+1}})$ for any integer $h > 0$. Recall that h corresponds to the number of hierarchical levels used in the scheme and that by increasing h, we get arbitrarily close to linear scaling. A natural question is whether there is a price to pay for this superior scaling of the throughput. In particular, what is the throughput-delay trade-off for hierarchical cooperation? In this chapter, we analyze the delay performance of the hierarchical cooperation scheme and show that the structure suggested in Section 4.4 is suboptimal from delay point of view. We devise a modification of the scheme in this chapter that achieves much better delay performance for the same throughput. More precisely, we show that one important drawback of the architecture in Section 4.4 is that it uses an extremely large bulk-size, where the bulk-size of a scheme refers to the minimum number of bits that should be communicated between each source-destination pair under this scheme. We show in Section 5.3.2 that the bulk-size used by the architecture in Section 4.4 scales like $B_h(n) = \Theta(n^{\frac{h}{2}})$; in other words, it grows arbitrarily fast as the throughput approaches linear scaling. Note that the bulk-size immediately imposes a lower bound on the end-to-end delay of each communication; even if there is no transmission delay from the source node to the destination node, receiving a bulk of $B(n)$ bits will take at least $\Theta(B(n)/\log n)$ channel uses for a destination node, since a simple application of the cut-set bound upper bounds the rate of reception by (or transmission from) a node with $\log n$ bits per channel use.

The basic idea behind the scheme in Section 4.4 is to first distribute the bits of a source node to its neighboring nodes, so that these bits can then be simultaneously transmitted to a group of nodes in the vicinity of the destination node. By collecting the observations of the receiving nodes to the actual destination node, the destination node is able to recover the bits intended for itself. The efficiency of these distributed MIMO transmissions increases with the size of (number of nodes contained in) the transmit and receive clusters, formed around the source node and the destination node respectively. However, when the size of the transmit cluster is large, the bulk of data that is communicated from each source node to its destination node has to be large as well. This bulk of data is to be chopped off and distributed among the many nodes in the transmit cluster. Hence, the size of the transmit cluster imposes

a lower bound on the bulk size that needs to be communicated between each source-destination pair. Moreover, in the hierarchical cooperation scheme, the cooperation traffic of distributing the information bits of the source node and collecting the MIMO observations to the destination node, is further handled by distributed MIMO communication. This resulting hierarchical architecture was shown to be efficient from throughput point of view. However, since distributed MIMO based communication imposes a lower bound on the bulk-size, repeating the idea recursively yields a scheme with even larger bulk-size. This is the reason why the bulk size of the hierarchical cooperation scheme increases as $\Theta(n^{\frac{h}{2}})$ with h hierarchical levels.

In this chapter, we suggest a modification of the hierarchical cooperation scheme that handles the problem of cooperation more efficiently. In order to do this, we study the problem of cooperation more carefully by posing it as a uniform traffic problem, instead of separating it into multiple permutation traffic problems, as was done in Section 4.4. In the uniform traffic problem, each of the n nodes in the network is interested in conveying independent information, say fixed L bits, to each of the other nodes in the network. In Section 5.3.2, we propose a two-phase hierarchical scheme that solves this uniform traffic problem in $\Theta(n^{\frac{h+1}{h}})$ time-slots, for any $h > 0$. Using this scheme for cooperation, the modified hierarchical cooperation scheme achieves the same aggregate throughput $T_h(n) = \Theta(n^{\frac{h}{h+1}})$ by using a much smaller bulk-size $B_h(n) = T_h(n)$. We show that reduced bulk size consequently reduces the delay and that the modified hierarchical cooperation scheme achieves $D_h(n) = \Theta(n)$.

We proceed by optimizing scheduling in this scheme to further reduce the end-to-end delay. To do this, we need to consider a generalized version of the uniform traffic problem where each node in the network is interested in conveying independent information to each of the nodes in a subset of $A(n)$ nodes, where the $A(n) < n$ nodes are chosen uniformly at random among the n nodes in the network. In Section 5.3.2, we show that this task can be accomplished in $\Theta(\frac{A(n)}{n} n^{\frac{h}{h+1}} \log n)$ channel uses for any $h > 0$, if $A(n) \geq n^{\frac{h}{h+1}}$. This allows us to achieve a throughput delay trade-off of $(T(n), D(n)) = (n^b / \log n, n^b \log n)$ for any $0 \leq b < 1$. This new result is depicted in Figure 5.1, together with previous results from the literature. In particular, the throughput-delay trade-off for the multi-hopping scheme has been established in [21, 22] as $D(n) = T(n)$ where $T(n)$ lies between $\Theta(1)$ and $\Theta(\sqrt{n})$. As shown in the figure, hierarchical cooperation complements this trade-off for $T(n)$ between $\Theta(\sqrt{n})$ and $\Theta(n)$.

A related line of research (see e.g. [21, 26, 39, 47]) is the characterization of the throughput-delay trade-off for mobile networks, where nodes move over the duration of communication according to a certain mobility pattern. In general, mobility schemes achieve an aggregate throughput scaling comparable to that of hierarchical cooperation (i.e. up to linear in n), but the delay scaling performance of such schemes may vary significantly, depending on the chosen mobility model. For instance, under the classical random walk mobility model considered in [21], the performance is quite poor, as illustrated in

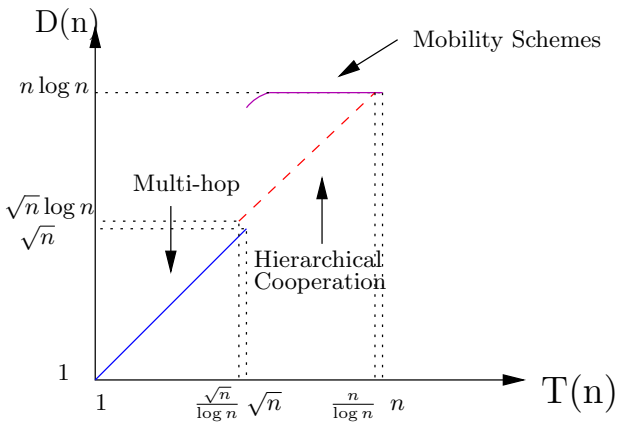

Figure 5.1: Throughput-delay performance achieved by hierarchical cooperation together with known results from the literature.

Figure 5.1. But from the delay point of view, a more prominent disadvantage which is common to all mobility models and which does not appear on the graph in Figure 5.1, is the constant that precedes the delay scaling law. Roughly speaking, this pre-constant relates to the speed of nodes in the case of mobility schemes, whereas it relates to the speed of light in the case of hierarchical cooperation or multi-hopping.

5.2 Setting and Main Results

The network and traffic model considered in this chapter is the one introduced in Section 2. Recall that the source and the destination nodes are paired up one-to-one and each source has the same traffic rate to send to its destination node. In this chapter, we will refer to this traffic pattern as permutation traffic in order to distinguish it from the uniform traffic problems that will be discussed in Sections 5.4 and 5.5.2. The aim of this chapter is to study the properties of the hierarchical cooperation scheme, so we assume the long-distance SNR in the network is large, $\mathrm{SNR}_l \gg 0$ dB, i.e the network operates in the degrees of freedom limited regime, Regime-I in (4.2).

Definition 5.2.1 (Delay of a Scheme)**.** *A scheme Π_n is said to achieve a delay D if all source destination pairs can communicate a message of $L > 1$ bits, in D time-slots under this scheme with probability $\lim_{n\to\infty} \mathbb{P}_{\Pi_n} = 1$ over the random realizations of the network.*

Note that so defined, the delay of a scheme quantifies the minimal time spent by the bits traveling inside the network while operated under this scheme. This definition of delay is consistent with [21, 22] and therefore, the comparison in Figure 5.1 of the multi-hop scheme and hierarchical cooperation is fair.

However, note that this definition does not include the queuing delay at the source node, as the clock starts when a packet leaves its source node. The delay at the source node can be accounted for by assuming a particular packet arrival process and studying the overall delay of a packet from its arrival at the source queue to the decoding at the destination node. The transmission delays given in Figure 5.1 can be regarded as lower bounds to this overall delay. Consider for example the simple TDMA scheme, one at a time transmission between the source-destination pairs, that corresponds to the origin in Figure 5.1. Assume independent Poisson packet arrival at each source node of appropriate rate. If we assume round-robin fashion, backlog unaware scheduling between the transmissions, the overall delay of the TDMA scheme will be $\Theta(n)$ much larger than the $\Theta(1)$ delay predicted by Figure 5.1. However, it is known that this delay can be reduced to $O(\log n)$ with backlog aware scheduling [38]. In general, how larger is the overall delay from the transmission delay given in Figure 5.1 depends on how well one can match the packet arrival process with backlog aware scheduling schemes. In this paper, our aim is to quantify the transmission delay of the discussed schemes; the second question regarding the queuing delay at the source is left open.

The following theorem is the main result of this chapter.

Theorem 5.2.1. *For $0 \leq b < 1$, the throughput-delay scaling*

$$(T(n), D(n)) = \Theta\left(n^b / \log n, n^b \log n\right)$$

is achievable using hierarchical cooperation. (See Figure 5.1.)

5.3 Delay of the Distributed MIMO Scheme with Hierarchical Cooperation

In this section, we establish the throughput-delay trade-off for the hierarchical cooperation scheme presented in Section 4.4. This requires a re-discussion of the scheme, but now with an emphasis on its delay performance. Below, we briefly summarize the main properties of the scheme when necessary, but the technical details earlier proven in Section 4.4.1 are mostly omitted.

Recall that the hierarchical cooperation scheme is based on clustering the nodes in the network and performing long-range MIMO transmissions between the clusters. The long-range MIMO transmissions should be preceded and followed by cooperation phases establishing transmit and receive cooperation respectively, which yields three successive phases in the operation of the network. If simple TDMA is used for establishing cooperation in phase 1 and 3, the overall scheme achieves a $\Theta(\sqrt{n})$-scaling of the aggregate throughput. This is the three phase scheme discussed in Section 5.3.1. Higher throughputs

can be achieved by setting the cooperation problem as multiple communication problems and using the three phase scheme as a solution to each of those communication problems. This yields the idea of recursion and results in a hierarchical architecture, where increasing the number of levels in the hierarchy yields an aggregate throughput scaling arbitrarily close to linear. The hierarchical cooperation scheme is discussed in detail in Section 5.3.2.

5.3.1 The Delay Scaling of the Three Phase Scheme

The network is divided into clusters of M_1 nodes and the scheme operates in three successive phases:

Phase 1: Setting Up Transmit Cooperation Clusters work in parallel. Within a cluster, each source node distributes LM_1 bits to the other nodes, L bits for each node where L is an arbitrarily large constant independent of M_1 and n, such that at the end of the phase, each node has L bits from each of the source nodes in its cluster. Since there can be at most M_1 source nodes in each cluster, this gives a traffic demand of exchanging at most $LM_1(M_1-1) < LM_1^2$ bits. Using TDMA, one-at-a-time transmission between pairs of nodes that has been shown in (4.8) to achieve a constant rate in each transmission, the L bits between each pair can be communicated in L/K_1 time-slots where $K_1 > 0$ is a constant independent of M_1 and n. Therefore, the LM_1^2 bits can be exchanged in a total of at most $(L/K_1)M_1^2$ time slots.

Phase 2: MIMO Transmissions Successive long-distance MIMO transmissions are performed between source-destination pairs, one at a time. In each one of the MIMO transmissions, say the one between s and d, the LM_1 bits of s are simultaneously transmitted by the M_1 nodes in its cluster to the M_1 nodes in the cluster of d. Each of the long-distance MIMO transmissions are repeated for each source-destination pair in the network, hence if the MIMO transmissions achieve a rate K_2M_1 for a constant $K_2 > 0$ independent of M_1, the phase can be completed in a total of $(L/K_2)n$ time-slots.

Phase 3: Cooperate to Decode Clusters work in parallel. Since there are at most M_1 destination nodes inside the clusters, each cluster received at most M_1 MIMO transmissions, each of length L/K_2 time-slots in phase 2. Each MIMO transmission is intended for a different destination node in the cluster. Thus, each node in the cluster has at most $(L/K_2)M_1$ received observations and each block of L/K_2 observations is to be conveyed to a different node in its cluster. Nodes quantize each observation into fixed Q bits, so there are a total of at most $(LQ/K_2)M_1^2$ bits to be exchanged inside each cluster. Using TDMA as in Phase 1, the phase can be completed in $(LQ/K_1K_2)M_1^2$ time slots.

In Section 4.4.1, it has been shown that each destination node is able to decode the transmitted bits from its source node from the quantized signals it gathers by the end of Phase 3. The throughput achieved by the scheme can be calculated as follows: each source node is able to transmit LM_1 bits to its destination node, hence LnM_1 bits in total are delivered to their destinations

in $(L/K_1)M_1^2 + (L/K_2)n + (LQ/K_1K_2)M_1^2$ time slots, yielding an aggregate throughput of

$$\frac{LnM_1}{(L/K_1)M_1^2 + (L/K_2)n + (LQ/K_1K_2)M_1^2}$$

bits per time-slot. Choosing $M_1 = \sqrt{n}$ to maximize this expression yields an aggregate throughput $T(n) = \frac{K_1K_2}{K_1+K_2+Q}\sqrt{n}$.

Note that as opposed to multi-hop, this three phase scheme allows only bulk transmission between any source-destination pair in the network; i.e. one can not arbitrarily communicate one bit (or L bits with L constant) using the three-phase scheme, but a minimum of $LM_1 = L\sqrt{n}$ bits should be transferred between every source-destination pair.

The end-to-end delay of this scheme is simply the total time for the three phases, since the bits are leaving their source nodes at the beginning of the first phase and are only decoded by their respective destination nodes at the end of the third phase. With the choice $M_1 = \sqrt{n}$, we see that the delay of the three phase scheme is $D(n) = ((L/K_1) + (L/K_2) + (LQ/K_1K_2))n$. Note that this delay scaling is much worse than the delay of the multi-hop scheme achieving same aggregate throughput.

5.3.2 The Hierarchical Cooperation Scheme

Higher aggregate throughput scaling can be achieved by using better network communication schemes than TDMA to establish the transmit and receive cooperations in the first and the third phases of the three phase scheme described in the previous section. Recall that there are LM_1^2 and $(LQ/K_2)M_1^2$ bits to be exchanged inside each cluster in phases 1 and 3, respectively. This traffic demand of exchanging LM_1^2 bits (or $(LQ/K_2)M_1^2$ bits) can be handled by setting up M_1 sub-phases, and assigning M_1 pairs in each sub-phase to communicate their L bits (or LQ/K_2 bits). The traffic to be handled at each sub-phase now looks similar to the original network communication problem (the permutation traffic defined in Section 5.2), with M_1 users instead of n. Any scheme suggesting a good solution for the original problem can now be used inside the sub-phases as an alternative to TDMA; for example, the multi-hop scheme and the three-phase scheme itself would be two alternatives both achieving an aggregate throughput scaling $\Theta(\sqrt{M_1})$ (in a network of size M_1) as opposed to the $\Theta(1)$ scaling achieved by TDMA.

Consider using the three phase scheme for cooperation as suggested in Section 4.4. More precisely, we want to handle the traffic of communicating L bit (or LQ/K_2 bits) between the M_1 pairs assigned in each sub-phase of phase 1 (or phase 3), by further dividing the clusters into smaller clusters of size M_2 and reusing the three phase scheme (TDMA-MIMO-TDMA). Note that this will create a hierarchical structure with two levels. See Figure 4.5. Note however that the three phase scheme in Section 5.3.1 allows only bulk transmissions between source-destination pairs. In this particular case, one will

have to communicate LM_2 bits between the source-destination pairs assigned at each sub-phase, as opposed to the original requirement of communicating only L bits (or LQ/K_2 bits). For the overall scheme, this in turn increases the bulk size to be communicated between every source-destination pair in the network from LM_1 bits to LM_1M_2 bits. So for the 2-level hierarchical scheme, we have to start by assuming that each source node in the network has LM_1M_2 bits to communicate to its destination node. It can be seen that these LM_1M_2 bits per source destination pair, or a total $n \times LM_1M_2$ bits in the network, can be communicated in

$$t_{total} = M_1 \left(\frac{L}{K_1} M_2^2 + \frac{L}{K_2} M_1 + \frac{LQ}{K_1 K_2} M_2^2 \right) + \frac{L}{K_2} M_2 n + M_1 \frac{Q}{K_2} \left(\frac{L}{K_1} M_2^2 + \frac{L}{K_2} M_1 + \frac{LQ}{K_1 K_2} M_2^2 \right) \tag{5.1}$$

time slots using the 2-level hierarchical scheme. The first term $M_1(L/K_1)M_2^2 + (L/K_2)M_1 + (LQ/K_1K_2)M_2^2)$ is the completion time of phase-1 of the three phase scheme. It is divided into M_1 sessions; in each session, M_1 source-destination pairs are assigned to communicate their LM_2 bits using a three phase scheme of TDMA-MIMO-TDMA. Recall from the computations of the three phase scheme in Section 5.2 that this takes $(L/K_1)M_2^2 + (L/K_2)M_1 + (LQ/K_1K_2)M_2^2$ time slots (M_1 and M_2 here correspond to the n and M_1, respectively, of the previous section). A similar argument holds for the third term $M_1(Q/K_2)((L/K_1)M_2^2 + (L/K_2)M_1 + (LQ/K_1K_2)M_2^2)$ in (5.1) which is the completion time for phase-3 with the extra Q/K_2 factor. Note that at the end of the first phase, each source node has distributed its LM_1M_2 bits among the M_1 nodes in its cluster, hence LM_2 bits for each node. These bits can be relayed to the destination cluster in $(L/K_2)M_2$ successive MIMO transmissions. Since the MIMO transmissions have to be repeated for each of the n source-destination pairs in the network, the completion time of the second phase is $(L/K_2)M_2n$ in (5.1).

Therefore, the aggregate throughput of the 2-level scheme is given by the expression

$$\frac{L\, M_2\, M_1\, n}{\frac{L}{K_2} M_2 n \;+\; M_1 \left(1 + \frac{Q}{K_2}\right) \left(\frac{L}{K_1} M_2^2 + \frac{L}{K_2} M_1 + \frac{LQ}{K_1 K_2} M_2^2 \right)} \tag{5.2}$$

and the optimal choices of $M_1 = n^{2/3}$ and $M_2 = n^{1/3}$ maximize the aggregate throughput scaling to

$$T_2(n) = \Theta(M_1) = \Theta\left(n^{2/3}\right),$$

while the denominator dictating the delay of the scheme is of order

$$D_2(n) = \Theta(M_2 \times n) = \Theta\left(n^{4/3}\right).$$

Note that with the 2-level hierarchical scheme, we improve the aggregate throughput scaling from $\Theta(\sqrt{n})$ for the three-phase scheme in the previous section to $\Theta\left(n^{2/3}\right)$, at the cost of increasing the bulk-size from $\Theta\left(\sqrt{n}\right)$ to $\Theta(n)$, which, in turn, increases the delay from $\Theta(n)$ to $\Theta\left(n^{4/3}\right)$.

The argument can be applied recursively to build an h-level hierarchical scheme. The optimal cluster size at the k'th level of an h-level hierarchical scheme can be computed as $M_k = \Theta\left(n^{\frac{h+1-k}{h+1}}\right)$. The aggregate throughput achieved by an h-level hierarchical cooperation scheme is given by

$$T_h(n) = \Theta(M_1) = \Theta\left(n^{\frac{h}{h+1}}\right).$$

The bulk-size is

$$B_h(n) = \Theta(M_h \times \ldots \times M_1) = \Theta\left(n^{\frac{h}{2}}\right)$$

and the end-to-end delay is

$$D_h(n) = \Theta(M_h \times M_{h-1} \times \cdots \times M_2 \times n) = \Theta\left(n^{\frac{h^2+h+2}{2(h+1)}}\right)$$

where we observe that for large h, the delay exponent is linear in h.

The results obtained in this section establish the poor delay performance of hierarchical cooperation. Note that the delay is mostly due to the large bulk-size used by the scheme. This is different from multi-hop schemes, since their bulk-size is constant ($\Theta(1)$) independent of the throughput achieved. The delay in multi-hop is rather due to the time spent in relaying the messages inside the network. In the next section, we suggest a modification of the scheme, so that it achieves the same throughput using much smaller bulk-size.

5.4 Hierarchical Cooperation with Smaller Bulk-Size

In this section, we treat the problem of cooperation in the three phase scheme with more care. We start by defining the uniform traffic problem to be the following.

Definition 5.4.1 (The Uniform Traffic Problem)**.** *Consider the assumptions on the network and channel model given in Section 5.2. Let each node in the network be interested in communicating independent information to each of the other nodes in the network. In particular, let us assume that each node has an independent L bits message to send to each of the other nodes in the network and the quantity of interest is the smallest time $F(n)$ required to accomplish this task. This problem we refer to as the uniform traffic problem.*

Note that as opposed the permutation traffic problem defined in Section 2, the problem is described in terms of the number of bits to be communicated between each source-destination pair and not the rate of communication. The following theorem provides an achievable solution to this problem.

Theorem 5.4.1. *For any integer $h > 0$, the uniform traffic problem can be solved in*

$$F(n) \leq K_3 n^{\frac{h+1}{h}}$$

time-slots w.h.p., for some constant $K_3 > 0$ independent of n.

Proof of Theorem 5.4.1: Let us start by assuming that there exists a scheme that solves the uniform traffic problem in $F(n) = LK_4n^b$ time-slots with $b > 1$ and $K_4 > 0$ a constant independent of n. Note that one such scheme is simple TDMA that yields $b = 2$. Using this existing scheme, we will construct a new scheme that yields smaller $F(n)$.

As before, let us start by dividing the network into clusters of M nodes. Let us first focus on one specific cluster S and one node d located outside of this cluster. In particular, all nodes in S have L bits to send to d. These bits can be communicated to d in two steps:

(1) The nodes in S *simultaneously* transmit their L bit messages destined to d forming a distributed transmit antenna array for MIMO transmission. The nodes in the destination cluster where d lies, form a distributed receive antenna array for this MIMO transmission.

(2) Each node in the destination cluster obtains one observation from the MIMO transmission in the previous phase; it quantizes and ships this observation to d, which can do joint MIMO processing of all the observations and decode the LM transmitted bits from the nodes in S.

As a first step towards handling the whole network problem, note that these two steps should be accomplished between S and all other nodes in the network. This can again be done in two steps:

Phase 1: MIMO transmissions We perform successive long-distance MIMO transmissions between S and all other nodes in the network. In each of the MIMO transmissions, say between S and d, the M nodes in S are simultaneously transmitting the L bit messages they would like to communicate to d and the M nodes in the cluster where d lies are observing the MIMO transmission. The MIMO transmissions should be repeated for each node in the network, if an aggregate rate of K_2M is achieved in each transmission for a constant $K_2 > 0$ and independent of M, we need $(L/K_2)n$ time-slots to complete the phase.

Phase 2: Cooperate to decode Clusters work in parallel. Since there are M nodes inside each cluster, each cluster received M MIMO transmissions from S of length L/K_2 time-slots in the previous phase. Each MIMO transmission is intended for different node in the cluster. Thus, each node in the cluster

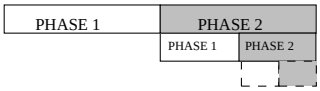

Figure 5.2: The figure illustrates the time-division in the hierarchical scheme that solves the uniform traffic problem.

has $(L/K_2)M$ observations, each block of L/K_2 observations is intended for a different node in the cluster. Each of these observations can be quantized into Q bits, with a fixed Q, which yields exactly the original uniform traffic problem, with M nodes instead of n and LQ/K_2 bits to be communicated between every pair of nodes. Using the scheme we assumed to exist in the beginning of the proof, this task can be completed in $(LQK_4/K_2)M^b$ time slots.

The total time we have spent during the two phases for handling the traffic originated from cluster S is given by $\frac{L}{K_2}n + \frac{LQK_1}{K_4}M^b$. From the network point of view, the above two steps should be completed for all n/M clusters in the network. Thus, the whole task can be completed in $\frac{n}{M}\left(\frac{L}{K_2}n + \frac{LQK_4}{K_2}M^b\right)$ time slots in total. Choosing $M = n^{\frac{1}{b}}$ in order to minimize this quantity yields

$$F(n) = L\,\frac{(1+QK_4)}{K_2}\,n^{2-\frac{1}{b}}.$$

Note that $2 - \frac{1}{b} < b$ for $b > 1$. In other words, we have established a solution for the uniform traffic problem that is better than the one we started with. Indeed, the two phase scheme described above can be used recursively yielding a better scheme at each step of the recursion. In particular, starting with TDMA achieving $b = 2$ and applying the idea recursively h times, one gets a scheme that solves the uniform traffic problem in $\Theta(n^{\frac{h+1}{h}})$ time slots. The operation of this scheme is illustrated in Figure 5.2. □

The interest in the uniform traffic problem arises from the fact that it exactly models the required traffic for cooperation in the three phase scheme. Recall the communication requirement inside the clusters in Phase 1 and 3 described in Section 5.3.1. This communication requirement, equivalent to a uniform traffic problem, is handled using TDMA in the three phase scheme which has been seen to be suboptimal from throughput point of view in Section 5.3.1. In the hierarchical cooperation scheme described in Section 5.3.2, this uniform traffic problem is handled by decomposing it into a number of permutation traffic problems. The resultant scheme is optimal in terms of throughput, but not very satisfying in terms of bulk-size. By using the solution to the uniform traffic problem suggested in this section, one can modify the hierarchical cooperation scheme, so as to achieve the same throughput with smaller bulk-size and consequently smaller delay. The resultant modified hierarchical scheme is illustrated in Figure 5.3. Note that the gain is coming

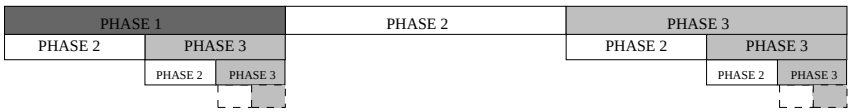

Figure 5.3: The figure illustrates the time-division in the modified hierarchical scheme that uses the scheme in Figure 5.2 for cooperation. Note the difference in operation of the phases between the modified hierarchical cooperation scheme and the original hierarchical cooperation scheme in Figure 4.5.

from treating the cooperation problem as it is and not necessarily as multiple permutation traffic problems as was previously done in Section 5.3.2.

Corollary 5.4.1. *A modified hierarchical cooperation scheme can achieve an aggregate throughput* $T_h(n) \geq K_5 n^{\frac{h}{h+1}}$ *with bulk-size* $B_h(n) = K_6 n^{\frac{h}{h+1}}$ *and delay* $D_h(n) \leq K_7 n$ *w.h.p., for any integer* $h \geq 0$ *and some positive constants* K_5, K_6, K_7 *independent of* n.

Proof of Corollary 5.4.1: Consider the three phase hierarchical scheme described in Section 5.3.1. By Theorem 5.4.1, the required traffic for transmit and receive cooperation in phase 1 and phase 3 can be handled in $LK_3M^{\frac{h+1}{h}}$ and $(LQ/K_2)K_3M^{\frac{h+1}{h}}$ time slots respectively. The expression for the aggregate throughput then becomes

$$\frac{LMn}{LK_3M^{\frac{h+1}{h}} + (L/K_2)n + (LQ/K_2)K_3M^{\frac{h+1}{h}}}$$

which is maximized by the choice $M = n^{\frac{h}{h+1}}$, yielding aggregate throughput $T_h(n) = \frac{K_2}{K_2K_3+1+K_3Q} n^{\frac{h}{h+1}}$, bulk-size $B_h(n) = LM = Ln^{\frac{h}{h+1}}$ and delay $D_h(n) = L(K_3 + 1/K_2 + K_3Q/K_2)n$. □

5.5 Hierarchical Cooperation with Better Scheduling

In the previous section, we presented a modified hierarchical scheme that achieves throughput $T_h(n) = \Theta(n^{\frac{h}{h+1}})$ using bulk-size $B_h(n) = \Theta(n^{\frac{h}{h+1}})$. However, the delay of this scheme is still $D_h(n) = \Theta(n)$. In this section, we optimize the scheduling in the scheme to further improve the delay performance to $D_h(n) = \Theta(n^{\frac{h}{h+1}} \log n)$. We first start by improving the scheduling in the three phase scheme with $h = 1$ discussed in Section 5.3.1. We then consider the modified hierarchical scheme with $h \geq 2$ discussed in Section 5.4 .

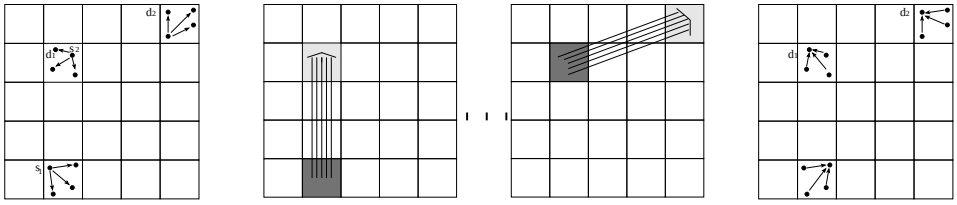

Figure 5.4: The three phase scheme with better scheduling. The figure illustrates the operation in one session.

5.5.1 Better Scheduling for the Three Phase Scheme

Recall the operation of the three phase scheme from the point of view of a single source-destination pair s-d as described in Section 5.3.1: a step (1) where s distributes its LM bits among the M nodes in its cluster, followed by a step (2) where these LM bits are simultaneously transmitted to the destination cluster via MIMO transmission, and a step (3) where the quantized MIMO observations are collected at the destination node d. These three steps need to be eventually accomplished for each source-destination pair in the network. In this section, we improve the scheduling in accomplishing this task: we organize M successive sessions and allow only n/M source-destination pairs to complete the three steps in each session.

In the beginning of each session we randomly choose one source node from each cluster, thus n/M source nodes in total. In general, the n/M destination nodes corresponding to these randomly chosen source nodes can be located anywhere. However, from Lemma 2.3.1-(c), we know that no more than $\log n$ of these destination nodes are located in the same cluster with high probability. We proceed by accomplishing the three steps for these chosen source-destination pairs:

Phase 1: Setting Up Transmit Cooperation Clusters work in parallel. The chosen source node in each cluster distributes its LM bits to the other nodes by using TDMA, which takes LM/K_1 time-slots in total if TDMA transmissions achieve a constant rate $K_1 > 0$. Note that as opposed to the scheme described in Section 5.3.1, there is only one source node operating in each cluster.

Phase 2: MIMO Transmissions Successive MIMO transmissions originated from each cluster are performed, transmitting the bits of the active source node in each cluster to its respective destination cluster. Note that in the current case, there is only one MIMO transmission originated from each cluster, so there are only n/M MIMO transmissions that need to be performed in total, each of length L/K_2 time slots if the MIMO transmissions achieve an aggregate rate K_2M for a constant $K_2 > 0$. This will require total time $(L/K_2)n/M$.

Phase 3: Cooperate to Decode Clusters work in parallel. Each cluster

received at most $\log n$ MIMO transmissions in phase 2 by Lemma 2.3.1-(c) and each MIMO transmission intended for a different destination node in the cluster. The received observations at each node are quantized into Q bits and are to be conveyed to the actual destination nodes. The traffic inside each cluster is at most of exchanging $(LQ/K_2)M \log n$ bits and can be completed using TDMA in at most $(LQ/K_1K_2)M \log n$ time slots. (See Figure 5.4.)

The operation continues with the next session by choosing a new set of n/M source nodes randomly among the nodes that have not yet accomplished the above steps. Note that all source-destination pairs will accomplish the three steps in a total of M sessions.

With this rather smoother operation on the network level, we accomplish to serve n/M source-destination pairs in each session, that is transfer $LM \times \frac{n}{M}$ bits in total to their destinations in $(L/K_1)M+(L/K_2)\frac{n}{M}+(LQ/K_1K_2)M \log n$ time slots yielding aggregate throughput

$$\frac{LM \times \frac{n}{M}}{(L/K_1)M + (L/K_2)\frac{n}{M} + (LQ/K_1K_2)M \log n} \tag{5.3}$$

which is maximized by the choice $M = \sqrt{n}$ yielding aggregate throughput $T(n) = \frac{K_1K_2}{K_2+K_1+Q}\frac{\sqrt{n}}{\log n}$. The delay experienced by each bit is now much less compared to the three phase scheme in Section 5.3.1, since it is again dictated by the total time spent in the three phases (denominator of (5.3)), which is now less than $D(n) = L(\frac{1}{K_1} + \frac{1}{K_2} + \frac{1}{K_1K_2})\sqrt{n} \log n$.

Note that instead of choosing $M = \sqrt{n}$, which is the optimal choice to maximize the throughput achieved by the scheme, one can choose $M = n^b$ with $0 \leq b \leq 1/2$. In this case, we also restrict the number of source-destination pairs to be served in each session to M, which can be less than the total number of clusters n/M. Indeed, we operate one source node in each of the $M(\leq n/M)$ clusters and simply keep the remaining clusters inactive. The expression for the aggregate throughput becomes

$$\frac{LM \times M}{(L/K_1)M + (L/K_2)M + (LQ/K_1K_2)M \log n}$$

which implies that the scheme achieves aggregate throughput $T(n) = \Theta(n^b/\log n)$ and delay $D(n) = \Theta(n^b \log n)$ for any $0 \leq b \leq 1/2$. Note that this throughput-delay trade-off differs only by $\log n$ from the trade-off achieved by multi-hop schemes.

5.5.2 Better Scheduling for the Hierarchical Cooperation Scheme

In this section, we adopt the scheduling idea of Section 5.5.1 to the modified hierarchical scheme presented in Section 5.4. However, this modification is not trivial and requires us to consider a generalized version of the uniform traffic problem.

Definition 5.5.1 (The Generalized Uniform Traffic Problem)**.** *Consider the assumptions on the network and channel model given in Section 5.2. Let each of the n nodes in the network be interested in conveying independent information to a subset $A(n)$ of the nodes ($A(n) \leq n$), where the $A(n)$ nodes are chosen randomly among the n nodes in the network. In particular, let us assume that each node in the network has an independent L bits message to send to each of these $A(n)$ nodes and the quantity of interest is the minimal time $G(n)$ required to accomplish this task. We define this to be the generalized uniform traffic problem.*

The following theorem provides an achievable solution to this problem. We skip the proof of the theorem since it is similar in spirit to the proof of Theorem 5.4.1.

Theorem 5.5.1. *For any integer $h > 0$, if $A(n) \geq n^{\frac{h}{h+1}}$, then the generalized uniform traffic problem can be solved in*

$$G(n) \leq LK_8 \frac{A(n)}{n} n^{\frac{h+1}{h}} \log(n)$$

time-slots w.h.p., for some constant $K_8 > 0$ independent of n.

Note that the generalized uniform traffic problem contains the uniform traffic problem discussed earlier as a special case with $A(n) = n$. Plugging $A(n) = n$ in Theorem 5.5.1, we recover the result of Theorem 5.4.1 with an extra $\log n$ factor. Indeed, when the condition $A(n) \geq n^{\frac{h}{h+1}}$ is satisfied with strict inequality in order, the extra $\log n$ factor in Theorem 5.5.1 is not needed. However, it is needed to account for the case $A(n) = n^{\frac{h}{h+1}}$, in which case it arises due to part-b of Lemma 2.3.1-(c).

We are now ready to apply the scheduling idea in Section 5.5.1 to the hierarchical cooperation scheme. Consider dividing the network into clusters of M_1 nodes and then further divide these clusters into smaller clusters of size M_2. Following the scheduling idea in Section 5.5.1, let us organize M_1/M_2 sessions and for each session randomly choose one small cluster inside every large cluster. Only the source nodes located in the chosen small clusters and their corresponding destination nodes will be served at each session. As usual, we are operating in three successive phases in each session:

Phase 1: Setting Up Transmit Cooperation The active small clusters operate in parallel. Note that there is a single active cluster of size M_2 inside every large cluster of size M_1. Let S_2 be the chosen small cluster inside the larger cluster S_1 that will operate in the current session. In this phase, each of the M_2 source nodes in S_2 need to distribute their LM_1 bits among the M_1 nodes in the larger cluster S_1, each block of L bits goes to a different node. This can be accomplished in two sub-phases:

- **Sub-Phase 1: MIMO transmissions** Successive MIMO transmissions are performed between nodes in S_2 and each node in S_1. In each of these

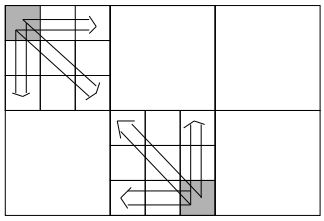

Figure 5.5: The figure illustrates sub-phase 1 of phase 1 of the modified hierarchical scheme with better scheduling. Note that there is only one small cluster distributing bits inside every large cluster.

MIMO transmissions, say the one between S_2 and a node d in S_1 (located outside of S_2), the M_2 nodes in S_2 are simultaneously transmitting the L-bits messages they would like to communicate to d. The M_2 nodes located in the same small cluster with d are acting as a distributed receive antenna array for this MIMO transmission. Since these MIMO transmissions should be repeated for every node in S_1, this sub-phase takes a total of $(L/K_2)M_1$ time-slots, if the MIMO transmissions achieve an aggregate rate K_2M for a constant $K_2 > 0$. See Figure 5.5.

- **Sub-Phase 2: Cooperate to Decode** All small clusters in the network work in parallel. In particular, each small cluster in S_1 has received M_2 MIMO transmissions from S in the previous phase, one MIMO transmission for each node in this small cluster. Thus, each node in the small cluster has $(L/K_2)M_2$ observations, one from each of the MIMO transmissions and each observation is to be conveyed to a different node in the cluster. Quantizing each observation into Q bits, we get the uniform traffic problem defined in Section 5.4 in a network of size M_2, and by Theorem 5.4.1 this problem can be handled in $(LQ/K_2)K_3M_2^{\frac{h_1+1}{h_1}}$ time-slots for any integer $h_1 > 0$.

Phase 2: MIMO Transmissions At the end of the first phase, all source nodes in the active small clusters have distributed their LM_1 bits among the nodes in the larger cluster. Now, successive long-distance $M_1 \times M_1$ MIMO transmissions between large clusters are performed. During each MIMO transmission, the LM_1 bits of a particular source node in the active small cluster are transferred to the destination cluster where its destination node is located. The number of MIMO transmissions to be performed in this phase is equal to the total number of source nodes active in this session. Hence the total phase can be completed in $(L/K_2)\frac{n}{M_1} \times M_2$ time-slots.

Phase 3: Cooperate to Decode By part-a of Lemma 2.3.1-(e), there are order M_2 destination nodes located in each of the large clusters. Thus, each large cluster has received M_2 MIMO transmissions in the previous phase, and the quantized MIMO observations spread over the M_1 nodes of the large

cluster should be collected at the corresponding M_2 destination nodes. This is the generalized uniform traffic problem of size M_1 with $A(M_1) = M_2$. By Theorem 5.5.1, it can be solved in $(LQ/K_2)K_8\frac{M_2}{M_1} \times M_1^{\frac{h_2+1}{h_2}} \log M_1$ time-slots for any integer $h_2 > 0$ provided that $A(M_1) \geq M_1^{\frac{h_2}{h_2+1}}$.

Gathering everything together, at every session of this modified hierarchical cooperation scheme, we deliver $LM_1 \times M_2 \times \frac{n}{M_1}$ bits to their destinations in a total of

$$\left(\frac{L}{K_2}M_1 + \frac{LQK_3}{K_2}M_2^{\frac{h_1+1}{h_1}}\right) + \frac{L}{K_2}\frac{n}{M_1} \times M_2 + \frac{LQK_8}{K_2}\frac{M_2}{M_1} \times M_1^{\frac{h_2+1}{h_2}} \log M_1$$

time-slots. The aggregate throughput is given by

$$\frac{\frac{n}{M_1} \times M_2 \times M_1}{\frac{L}{K_2}M_1 + \frac{LQK_3}{K_2}M_2^{\frac{h_1+1}{h_1}} + \frac{L}{K_2}\frac{n}{M_1} \times M_2 + \frac{LQK_8}{K_2}\frac{M_2}{M_1} \times M_1^{\frac{h_2+1}{h_2}} \log M_1}$$

which is maximized by the choice $h = h_2 = h_1 + 1$, $M_1 = n^{\frac{h}{h+1}}$ and $M_2 = M_1^{\frac{h-1}{h}}$, yielding aggregate throughput $T(n) = \Theta\left(\frac{n^{\frac{h}{h+1}}}{\log n}\right)$ and delay $D(n) = \Theta\left(n^{\frac{h}{h+1}} \log n\right)$. Note that these choices for M_1 and M_2 satisfy the constraint $A(M_1) = M_2 \geq M_1^{\frac{h_2}{h_2+1}}$.

Note that at this point, we have proven that all points on the throughput-delay scaling curve $(T(n), D(n)) = \left(\Theta\left(n^{\frac{h}{h+1}}/\log n\right), \Theta\left(n^{\frac{h}{h+1}} \log n\right)\right)$ with h being a positive integer are achievable. In order to show that all points on the line $(T(n), D(n)) = (\Theta(n^b/\log n), \Theta(n^b \log n))$ with $0 \leq b < 1$ are achievable, we can choose $M_1 = n^b$ with $0 \leq b \leq \frac{h}{h+1}$ in the above discussion, while maintaining the relationships $M_2 = M_1^{\frac{h-1}{h}}$ and $h = h_2 = h_1 + 1$. Extending the argument at the end of Section 5.5.1, we also restrict the number of small clusters to be served in each session to $M_1^{1/h}$ which can now be less than the total number of large clusters n/M_1 ($\geq M_1^{1/h}$). Indeed, we operate one small cluster in each of the $M_1^{1/h}$ large clusters and simply keep the remaining large clusters inactive. The expression for the aggregate throughput becomes

$$\frac{LM_1^{\frac{1}{h}} \times M_2 \times M_1}{\frac{L}{K_2}M_1 + \frac{LQK_3}{K_2}M_2^{\frac{h_1+1}{h_1}} + \frac{L}{K_2}M_1^{\frac{1}{h}} \times M_2 + \frac{LQK_8}{K_2}\frac{M_2}{M_1} \times M_1^{\frac{h_2+1}{h_2}} \log M_1}$$

which shows that we can achieve aggregate throughput $T(n) = \Theta(M_1/\log M_1)$ and delay $D(n) = \Theta(M_1 \log M_1)$. Recalling that $M_1 = n^b$, we get the points on the throughput-delay scaling curve $(T(n), D(n)) = (n^b/\log n, n^b \log n)$ for any $0 \leq b \leq \frac{h}{h+1}$ and $h > 0$. This concludes the proof of the main result of this chapter. □

6 Outlook

The results and the conclusions of the thesis have already been summarized in the introduction. In this chapter, we want to discuss how the results connect with and compare to some of the concurrent literature. We also review follow-up works that extend some of the ideas in this dissertation. The discussion points toward some new research directions and open problems.

6.1 Other Operating Regimes of Wireless Networks

Are there any other operating regimes in large wireless networks than those discussed here? A complete answer to this question is out of the scope of the current dissertation. However, the scaling law formulation developed in this thesis for identifying the operating regimes of wireless networks is general, and can be extended to study such problems. New operating regimes can be discovered by diversifying the assumptions that have led to the four operating regimes in this thesis. The current section explores two such assumptions. The discussion suggests two new operating regimes for wireless networks and demonstrates how some recent works in the literature fit in the framework suggested here.

The first assumption that we discuss is the power law coupling between the system resources that has been assumed in Section 1.2 and later in Chapter 2 when formulating the scaling law problem. Recall that the scaling exponent of the spectral efficiency has been characterized when $\mathrm{SNR}_s = n^\beta$ for some real and finite β, when a priori, there are other ways to couple SNR_s and n. The power law relation is the one that is naturally suggested by the inherent spatial structure of the problem and the power decay law with distance. As

SNR_l and SNR_s have a ratio of $n^{1-\alpha/2}$, this kind of scaling appears to lead to the richest possible classification. This is also evidenced by the discovery of four qualitatively different operating regimes. This is analogously the case for the AWGN channel. The power law coupling $\mathrm{SNR} = m^\gamma$ for any real γ leads to the discovery of the two operating regimes of the AWGN channel. (See Section 1.2.) However, note also the limitations of this formulation. In the first regime when $\mathrm{SNR} \gg 0$ dB, the dependence of the AWGN capacity on its two resources is approximated as $C(W, P_r/N_0) \propto W$, when actually there is also the logarithmic dependence on P_r/N_0W. The scaling law formulation with power law coupling $\mathrm{SNR} = m^\gamma$ misses out the logarithmic dependence on SNR. The resultant approximation is reasonable in the sense that the linear dependence on W is more important than the logarithmic dependence on P_r/N_0W. However, when P_r/N_0W is extremely large, the capacity is better approximated by $C(W, P_r/N_0) \propto W \log(P_r/N_0W)$. This approximation captures the fact that the capacity grows to infinity with increasing SNR. This regime can be discovered by studying the scaling law problem for $\mathrm{SNR} = e^{m^\gamma}$, for any $\gamma > 0$. The exponential coupling emphasizes the logarithmic term.

What happens in the case of wireless networks when SNR_s and n are coupled as $\mathrm{SNR}_s = e^{n^\beta}$ for $\beta > 0$? In this case, the spatial differentiation between nodes is lost. The connections between nodes are so strong that effectively, all the links in the network are of identical strength. Since spatial reuse is based on the fact that connections between far away nodes are so weak that the interference from such nodes will be close to the noise level, it can be expected that any scheme that is based on spatial reuse will not be efficient in this regime. Indeed, it is easy to observe that in this regime, even the scaling performance of simple TDMA between source-destination pairs can outperform the spatial reuse based schemes discussed in this thesis, such as hierarchical cooperation and multi-hopping. Optimal operation can be achieved by schemes that achieve linear scaling without making use of spatial reuse. The interference alignment scheme suggested recently in [9] can be investigated for optimality in this regime. We present a first attempt in this direction in [43]. See also [37].

A second assumption that can be questioned is the independence assumption of the phases in the channel model. Recall that the channel introduces a path loss attenuation to the transmitted signals and a random rotation in their phase. The phase rotations are assumed to be independent and identically distributed across pairs of nodes in the network. The motivation in studying this channel model is two-fold: first of all, this is a well-established model, used in almost all works in wireless communication and all the earlier works on scaling laws. Second, we anticipate it to be the relevant model for most wireless network applications in practice. However, when the nodes are packed together in a small area, the phases between different pairs of nodes become correlated and the independence assumption breaks down. This phenomenon has been recently studied in [19]. The work [19] shows that the number of spatial degrees of freedom in the network are upper bounded by

$\sqrt{A}/\lambda_c$, where λ_c is the carrier wavelength of the signals. Equivalently, the total number of degrees of freedom per Hz in the network is upper bounded by the minimum of the number of users n and $\sqrt{A}/\lambda_c$. This roughly implies that the conclusions based on the i.i.d. random phase model hold only if $A \geq \lambda_c n^2$. Note that this condition could be imposed on the scaling law formulation in Section 1.2 and would lead to the same four operating regimes, but with the extra condition that $A \geq \lambda_c n^2$. In that case, we would study the interplay between $A = n^{\beta_1}$, $P = n^{\beta_2}$ and $W = n^{\beta_3}$ for any real β_2, β_3 but $\beta_1 \geq 2$. Note that this again leads to $\mathrm{SNR}_s = n^{\beta}$ for any real β. (It can be readily verified from the definition of SNR_s in (1.3) that $\beta = \beta_2 - \beta_3 + \alpha(1 - \beta_1)/2$.)

From a practical point of view, the result of [19] implies that for networks with $\sqrt{A}/\lambda_c \geq n$, the random phase model holds and thus, such networks are characterized by one of the four operating regimes identified in this dissertation. For an area of $1\mathrm{km}^2$ and a carrier frequency of 3 GHz, this roughly implies that up to 10000 users can be accommodated in the network without experiencing any spatial degrees of freedom limitation. When $\sqrt{A}/\lambda_c < n$, the spatial degrees of freedom limitation comes into play. It can be expected that such networks still benefit from distributed MIMO communication when $\sqrt{n} < \sqrt{A}/\lambda_c < n$. For the numerical example above, this corresponds to up to 100 000 000 users in an area of $1\mathrm{km}^2$. When $\sqrt{A}/\lambda_c \leq \sqrt{n}$, it is expected that there is no need to go beyond multi-hopping as distributed MIMO communication can not provide any further benefit. The formulation proposed in this thesis can be extended to explore this spatial degrees of freedom limited regime.

6.2 Improving the Performance of the New Schemes

The new schemes presented in this dissertation are only investigated from a throughput scaling point of view, with the aim to demonstrate their potentials in large wireless networks. The only exception is Chapter 5, where the hierarchical cooperation scheme has been analyzed from a joint throughput-delay point of view. As demonstrated in that chapter, although the schemes presented in this dissertation achieve optimal throughput scaling, they can be improved in many other aspects.

One interesting research direction is to estimate the constants preceding the scaling laws. The work [24] presents a first step in this direction by providing a closer look to the pre-constant of the hierarchical cooperation scheme in Section 4.4. Note that in Theorem 4.2.1, both the pre-constant and the scaling exponent depend on ϵ. This dependence comes from the fact that both the scaling exponent and the pre-constant depend on the number of hierarchical levels in the scheme. This dependence is analyzed in [24] and it is shown that there is no benefit in increasing the number of hierarchical levels beyond

$\sqrt{\log n}$ in a network with a finite number of users n.

When it comes to the design and performance analysis of the schemes for a network with a given number of users, there are many optimization possibilities. The scaling law results in this thesis provide some architectural guidelines on how to design schemes that scale well. However, a detailed design and performance analysis would involve the tuning of many parameters and improvements of the schemes in order to optimize the pre-constant in the system throughput. For example, the schemes considered in this dissertation, both hierarchical cooperation and multi-hopping, make use of spatial reuse. Many local clusters operate simultaneously inside the network and a guard zone is imposed around each active cluster to control inter-cluster interference. The size of the guard zone is a design parameter to be optimized. Choosing a large guard zone decreases the resultant inter-cluster interference power, but leads to a larger overhead due to TDMA between neighboring clusters. This trade-off dictates the pre-constant of the throughput, but does not change the scaling law. Another example is the quantization of the received analog signals in the third phase of the hierarchical cooperation scheme. Each node in the received cluster quantizes the signal it observes and forwards the bits to the final destination. However, the quantized signals are correlated across the receive nodes. Hence, a reduction in the overhead can be achieved by doing some Wyner–Ziv coding (see [31], [44]).

Another research direction is to extend the results and ideas in this thesis to other settings with different traffic patterns and network models. The uniform traffic problem and the generalized uniform traffic problem considered in Chapter 5, are two such examples. In this line, the work [41] considers a more general setup by allowing the traffic demand between the nodes in the network to be arbitrary. In [40], the authors extend the results in this thesis to the case where the users are placed arbitrarily in the network area. Recall that in this thesis, we consider random placement of nodes which yields highly regular configurations (see Lemma 2.3.1). The extension to the arbitrary case requires an interesting modification of the hierarchical cooperation scheme. The distributed MIMO transmissions are replaced by distributed MIMO-MAC transmissions followed by distributed MIMO-BC transmissions. The cooperation required between the MAC and BC transmissions is achieved in a hierarchical fashion. The authors also show that the arbitrary placement of nodes may require to use a combination of hierarchical cooperation and multi-hopping in the fourth regime in (4.2). Recall that in the case of random networks, pure multi-hopping is sufficient to achieve optimal capacity scaling in this regime.

A number of open problems remain beyond the scope of the current thesis. One interesting problem is how to acquire the necessary channel state information for the new schemes in this thesis and to estimate the associated overhead. Recall that we have assumed full channel state information at all the nodes in the network. In fact, our schemes require a slightly less optimistic assumption. The channel state information is needed only at the receiver side. Nevertheless, in order for the destination nodes to be able to decode their

messages, they need to know the associated MIMO matrix together with the quantized MIMO observations of their cluster. In other words, nodes need to know not only the states of the channels to themselves but the channel states to a group of nodes around them. This may introduce significant overhead to communication when the channels are changing rapidly. Similar open problems are the estimation of the overhead required for cluster formation and the coordination of the network in a decentralized fashion.

Bibliography

[1] S. Aeron and V. Saligrama, "Wireless ad hoc networks: Strategies and scaling laws for the fixed snr regime," *IEEE Trans. Inform. Theory*, vol. 53, no. 6, pp. 2044–2059, 2007.

[2] R. Ahlswede, "Multi-way communication channels," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 279–280, 1974.

[3] S. Ahmad, A. Jovicic, and P. Viswanath, "Outer bounds to the capacity region of wireless networks," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2770–2776, 2006.

[4] A. Avestimehr, S. Diggavi, and D. Tse, "A deterministic approach to wireless relay networks," in *Proceedings of Allerton Conference on Communication, Control, and Computing*, Illinois, USA, 2007.

[5] ——, "Wireless network information flow," in *Proceedings of Allerton Conference on Communication, Control, and Computing*, Illinois, USA, 2007.

[6] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 19, no. 2, pp. 197–207, 1973.

[7] ——, "A simple converse for broadcast channels with additive white gaussian noise," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 279–280, 1974.

[8] H. Bolcskei, R. Nabar, O. Oyman, and A. Paulraj, "Capacity scaling laws in mimo relay networks," *IEEE Trans. Inform. Theory*, vol. 5, no. 6, pp. 1433–1444, 2006.

[9] V. Cadambe and S. Jafar, "Interference alignment and the degrees of freedom for the k user interference channel," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3425–344, 2008.

[10] T. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[11] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inform. Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[12] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[13] A. Dana and B. Hassibi, "On the power efficiency of sensory and ad-hoc wireless networks," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, pp. 2890–2914, 2006.

[14] S. Diggavi, M. Grossglauser, and D. Tse, "Even one-dimensional mobility increases the capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3947–3854, 2005.

[15] R. Etkin and D. Tse, "Degrees of freedom in some underspread mimo fading channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1576–1608, 2006.

[16] R. Etkin, D. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inform. Theory*, vol. 54, no. 12, pp. 5534–5562, 2008.

[17] G. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell System Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[18] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inform. Theory*, vol. 53, no. 3, pp. 1009–1018, 2007.

[19] M. Franceschetti, M. Migliore, and P. Minero, "The capacity of wireless networks: information-theoretic and physical limits," 2007, pre-print.

[20] R. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[21] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks-part i: The fluid model," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2568–2592, 2006.

[22] ——, "Optimal throughput-delay scaling in wireless networks-part ii: Constant-size packets," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5111–5116, 2006.

[23] A. E. Gamal, M. Mohseni, and S. Zahedi, "Bounds on capacity and minimum energy-per-bit for awgn relay channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1545–1561, 2006.

[24] J. Ghaderi, L.-L. Xie, and X. Shen, "Throughput optimization for hierarchical cooperation in ad hoc networks," in *Proceedings of IEEE International Conference on Communications*, Beijing, China, 2008.

[25] R. Gowaikar, B. Hochwald, and B. Hassibi, "Communication over a wireless network with random connections," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, pp. 2857–2871, 2006.

[26] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 4, pp. 477–486, 2002.

[27] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[28] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inform. Theory*, vol. 27, no. 1, pp. 49–60, 1981.

[29] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[30] A. Jovicic, P. Viswanath, and S. Kulkarni, "Upper bounds to transport capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2555–2565, 2004.

[31] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.

[32] G. Kramer, I. Maric, and R. Yates, "Cooperative communications," *Foundations and Trends in Networking*, vol. 1, no. 3-4, pp. 271–425, 2007.

[33] S. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 1041–1049, 2004.

[34] O. Leveque and E. Telatar, "Information theoretic upper bounds on the capacity of large, extended ad-hoc wireless networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 858–865, 2005.

[35] H. Liao, "Multiple access channels," Ph.D. dissertation, Department of Electrical Engineering, University of Hawaii, Honolulu, 1972.

[36] V. Morgenshtern, H. Bolcskei, and R. Nabar, "Distributed orthogonalization in large interference relay networks," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Adelaide, Australia, 2005.

[37] B. Nazer, M. Gastpar, S. Jafar, and S. Vishwanath, "Ergodic interference alignment," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, 2009.

[38] M. Neely, E. Modiano, and Y. Cheng, "Logarithmic delay for $n \times n$ packet switches under the crossbar constraint," *IEEE/ACM Trans. on Networking*, vol. 15, no. 3, pp. 657–668, 2007.

[39] M. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1917–1937, 2005.

[40] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," 2007, eprint, arXiv: 0711.2745, arxiv.org.

[41] ——, "The unicast and multicast capacity regions of large wireless networks," 2008, eprint, arXiv: 0809.1344v2, arxiv.org.

[42] A. Ozgur, O. Leveque, and E. Preissmann, "Scaling laws for one and two-dimensional random wireless networks in the low attenuation regime," *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3573–3585, 2007.

[43] A. Ozgur and D. Tse, "Achieving linear scaling with interference alignment," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, 2009.

[44] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer, "Communication via decentralized processing," *IEEE Trans. Inform. Theory*, vol. 54, no. 7, pp. 3008–3023, 2008.

[45] C. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[46] ——, "Two-way communication channels," in *Proc. of the 4th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, Berkeley, USA, 1961, pp. 611–644.

[47] G. Sharma, R. Mazumdar, and N. Shroff, "Delay and capacity tradeoffs in mobile ad hoc networks: A global perspective," in *Proc. of IEEE Infocom*, Barcelona, Spain, 2006.

[48] R. Stanley, *Enumerative Combinatorics, Vol. 2.* Cambridge: Cambridge University Press, 1999.

[49] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. on Telecommunications*, vol. 10, no. 6, pp. 585–596, 1999.

[50] S. Toumpis and D. Toumpakaris, "Wireless ad hoc networks and related topologies: applications and research challenges," *Elektrotechnik und Informationstechnik, Springer*, vol. 123, no. 6, pp. 232–241, 2006.

[51] E. van der Meulen, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, no. 1, pp. 120–154, 1971.

[52] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[53] L.-L. Xie and P. R. Kumar, "On the path-loss attenuation regime for positive cost and linear scaling of transport capacity in wireless networks," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2313–2328, 2006.

[54] L.-L. Xie and P. Kumar, "A network information theory for wireless communications: Scaling laws and optimal operation," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 748–767, 2004.

Curriculum Vitae

Education:

- **Swiss Federal Institute of Technology at Lausanne (EPFL)**
  Ph.D. in Information Theory
  Research Topic: Wireless Adhoc Networks
  August 2005 - July 2009 (Expected).

  Pre-doctoral Certificate.
  School of Computer and Communication Sciences, EPFL.
  October 2004 - July 2005

- **Middle East Technical University, Ankara**
  M.Sc in Electrical and Electronics Engineering
  Specialization: Communication Systems
  September 2001 - July 2004.

  B.Sc in Electrical and Electronics Engineering
  September 1997 - June 2001

  B.Sc in Physics (Double Major Program)
  September 1997 - June 2001

Professional Experience:

- **Scientific and Technical Research Council of Turkey, Ankara**
  Defense Industries Research and Development Institute
  Full-Time Hardware Design Engineer
  August 2001 - August 2004

Publications:

Monographs:

- A. Özgür, O. Lévêque, D. Tse, *Operating Regimes of Wireless Networks*, in preparation for Foundations and Trends in Networking, Now Publishers.

Journal Publications:

- A. Özgür, R. Johari, D. Tse, O. Lévêque, *Information Theoretic Operating Regimes of Large Wireless Networks*, to appear in IEEE Transactions on Information Theory, submitted February 2008.
- A. Özgür, O. Lévêque, *Throughput-Delay Tradeoff for Hierarchical Cooperation in Ad Hoc Wireless Networks*, to appear in IEEE Transactions on Information Theory, submitted December 2007.
- A. Özgür, O. Lévêque, D. Tse, *Hierarchical Cooperation Achieves Optimal Capacity Scaling in Ad Hoc Networks*, IEEE Transactions on Information Theory 53 (10), October 2007, 3549–3572.
- A. Özgür, O. Lévêque and E. Preissmann, *Scaling Laws for One and Two-Dimensional Random Wireless Networks in the Low Attenuation Regime*, IEEE Transactions on Information Theory 53(10), October 2007, 3573–3585.

Conference Publications:

- A. Özgür, D. Tse, *Achieving Linear Scaling with Interference Alignment*, to be presented in IEEE International Symposium on Information Theory, Seoul, July 2009.
- A. Özgür, R. Johari, D. Tse, O. Lévêque, *Information Theoretic Operating Regimes of Large Wireless Networks*, Proc. IEEE Int. Symposium on Information Theory, Toronto, July 2008.
- A. Özgür, O. Lévêque, *Throughput-Delay Tradeoff for Hierarchical Cooperation in Ad Hoc Wireless Networks*, Proc. Int. Conference on Telecommunications, St Petersburg, June 2008.
- A. Özgür, O. Lévêque and D. Tse, *Exact Capacity Scaling of Extended Wireless Networks*, Proc. IEEE Int. Symposium on Information Theory, Nice, July 2007.

- A. Özgür, O. Lévêque and D. Tse, *Hierarchical Cooperation Achieves Linear Capacity Scaling in Ad Hoc Networks*, Proc. IEEE Infocom Conference, May 2007.
- A. Özgür, O. Lévêque and D. Tse, *How does the Information Capacity of Ad Hoc Networks Scale?*, invited paper at the Allerton Conference on Communication, Control and Computing, October 2006.
- A. Özgür and O. Lévêque, *Scaling Laws for Two-Dimensional Random Ad-Hoc Wireless Networks*, Proc. IEEE Int. Zurich Seminar on Communications, Zurich 2006.