Scaling Laws for Heterogeneous Wireless Networks

by

Urs Niesen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author .....
Department of Electrical Engineering and Computer Science
August 14, 2009

Certified by .....
Devavrat Shah
Associate Professor
Thesis Supervisor

Certified by .....
Gregory W. Wornell
Professor
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Scaling Laws for Heterogeneous Wireless Networks

by

Urs Niesen

Submitted to the Department of Electrical Engineering and Computer Science
on August 14, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis studies the problem of determining achievable rates in heterogeneous wireless networks. We analyze the impact of location, traffic, and service heterogeneity. Consider a wireless network with n nodes located in a square area of size n communicating with each other over Gaussian fading channels. Location heterogeneity is modeled by allowing the nodes in the wireless network to be deployed in an arbitrary manner on the square area instead of the usual random uniform node placement. For traffic heterogeneity, we analyze the $n \times n$ dimensional unicast capacity region. For service heterogeneity, we consider the impact of multicasting and caching. This gives rise to the $n \times 2^n$ dimensional multicast capacity region and the $2^n \times n$ dimensional caching capacity region. In each of these cases, we obtain an explicit information-theoretic characterization of the scaling of achievable rates by providing a converse and a matching (in the scaling sense) communication architecture.

Thesis Supervisor: Devavrat Shah
Title: Associate Professor

Thesis Supervisor: Gregory W. Wornell
Title: Professor

Acknowledgments

Many people have helped me with this thesis and throughout my time at MIT. First, I would like to thank my two advisors, Devavrat Shah and Greg Wornell. I had the good fortune to be advised by two individuals that truly cared about my development — both professional as well as personal — and I learned a lot from both of them. Further thanks go to Lizhong Zheng for agreeing to serve on my thesis committee.

I am also grateful to Piyush Gupta and Mitch Trott for hosting me during my internships at Bell Labs and HP Labs, respectively. Further thanks go to Uri Erez, Olivier Lévêque, Sanjoy Mitter, Aslan Tchamkerten, Emre Telatar, and David Tse for their influence during various stages of this research.

Thanks go also to the administrative staff of LIDS and SIA, and in particular Tricia O'Donnell and Lynne Dell. Their help throughout these years is much appreciated.

My time in Cambridge would not have been the same without my friends and colleagues at MIT. In particular, I would like to thank Anthony Accardi, Emmanuel Abbé, Shashi Borade, Venkat Chandar, Venkat Chandrasekaran, Vijay Divi, Vishal Doshi, Ying-zong Huang, Ashish Khisti, Minji Kim, Yuval Kochman, James Krieger, Evgeny Logvinov, Baris Nakiboglu, Mesrob Ohannessian, Parikshit Shah, Maryam Modir Shanechi, Charles Swannack, Katy Thorn, Kush Varshney, Lav Varshney, Da Wang. Finally, I would like to thank my parents and Preeti Kamakoti for their love and support when I needed it most.

This research was supported, in part, by the National Science Foundation under Grant No. CCF-0635191, by DARPA under Grant No. 18870740-37362-C, and by Hewlett-Packard under the MIT/HP Alliance. This funding is greatly appreciated.

Contents

Chapter 1

Introduction

Over the past decades, there has been a growing disconnect between the size of communication networks that are built and planned and the size of communication networks that are fundamentally understood. On the one hand, wireline networks (like the Internet) have grown from only a few hundred users in 1981 to over one billion in 2008, and wireless networks (like metropolitan mesh networks, sensor networks, or military ad-hoc networks) with up to a million communication devices are being envisioned. On the other hand, even simple communication networks, as for example a four node wireless network with two sources and two destinations, or an even simpler three node network with one source, one destination, and one relay, are only partially understood. Central questions remain unanswered: What is the role of interference in the four node example above, and what is the role of cooperation in the three node example? An answer to these questions will undoubtedly have profound implications on the design of future communication networks.

A main reason for this disconnect is that much of the effort analyzing these communication systems has been directed at obtaining exact solutions for small networks and trying to gain insight for larger networks from it. This has proved challenging, as the lack of a complete understanding of even very simple networks like the ones mentioned above illustrates. Another approach is to directly consider large networks, but instead settle for an approximate asymptotic solution.

To analyze such large networks, a model of how they are generated has to be

chosen. More precisely, consider a wireless network with n nodes. How should the location of these n nodes be chosen; how should the traffic demand they generate behave; and how should the services they require be modeled? This question is usually addressed by making several homogeneity assumptions. For the node locations, it is usually assumed that nodes are placed uniformly at random on a square of area n; for the traffic demands that each node is source for exactly one destination node chosen uniformly at random from among all the other nodes, and that all these n source-destination pairs communicate at equal rate; for the service requirements that all nodes generate only unicast traffic.

While this homogeneous setting is convenient mathematically, it does not provide a very accurate model of reality. In fact, for the node locations it is likely that some areas are denser than others (e.g., towns vs. countryside); for the traffic demands that users communicate to nearby nodes more often than to faraway ones, and that some users will create more traffic than others (e.g., sending an email vs. streaming a movie); for the service requirements that some information needs to be transmitted to several or all nodes. In other words, we expect node location, traffic demands, and service requirements to be highly heterogeneous. Moreover, these heterogeneities will lead to different asymptotic behavior of the network. This implies that the results obtained for large homogeneous wireless networks will only yield a limited understanding of the heterogeneous networks we are likely to encounter in practice.

In this thesis, we develop approximate asymptotic characterizations of the performance of large heterogeneous wireless networks. We consider the impact of location, traffic, and service heterogeneity. The common approach to deal with these heterogeneities consists of first finding the underlying "coarse structure" of the network, capturing the essential parts of the heterogeneity. Once such a simple coarse structure is identified, rather complicated questions about the network can be elegantly analyzed by recasting them for the underlying coarse structure. Moreover, this coarse structure allows to obtain insight into the role of interference or cooperation in large networks and can guide the design of communication schemes and algorithms.

1.1 Network Models

As mentioned in the previous section, to analyze large networks a model for their generation has to be chosen. Here we briefly review some popular such models used throughout the literature.

First, a model for the node location needs to be chosen. A standard assumption is that the n nodes of the wireless network are located on a square of area[1] n. It is often assumed that the nodes are placed uniformly and independently at random on this square, which we refer to as *random node placement*. If nodes are allowed to be placed in an arbitrary deterministic manner on this square, we speak of an *arbitrary node placement*. In the case of arbitrary node placement, it is usually assumed that there is some constant (independent of n) minimum separation between the nodes. This minimum-separation requirement prevents degenerate node placements.

Second, a model for communication between these nodes needs to be selected. There are two broad categories of such models. Models in the first category are motivated by current wireless technology and are referred to as *protocol models*. We describe two of them in more detail.

Protocol Model 1: Node v can receive data from node u at rate 1 bits/s if it lies outside the region of interference of each other transmitter.

Protocol Model 2: Node v can receive data from node u at rate $\log(1 + \text{SINR})$ bits/s, where SINR is the signal to interference plus noise ratio at the receiving node v. Here signals are attenuated as $r^{-\alpha/2}$ over distance r for some *path-loss exponent* $\alpha > 2$.

These communication models share two assumptions. First, they only allow point-to-point communication, and second, they treat all interference as noise. In other words, these models makes assumptions on the communication protocol used between these

[1]This is referred to as *extended node placement*. When nodes are located on a square of area 1 for any n, we speak of a *dense node placement*. Results for the two cases are closely related, and we focus on the extended case in this thesis.

nodes[2]. These two assumptions imply that the only allowed communication scheme in the wireless network is *multi-hop routing*, in which a message travels over multiple hops from its source to its destination, and each node along this route decodes the message received from the previous node and then re-encodes it for the subsequent one.

Models in the second category do not make any assumptions about the communication protocol used, but rather aim at directly describing the underlying wireless channel. Two popular models are the following.

Gaussian Model: Signals transmitted at node u are received at node v at distance r attenuated by a factor $r^{-\alpha/2}$ for some path-loss exponent $\alpha > 2$, and then further corrupted by additive Gaussian noise.

Gaussian Fading Model: Signals transmitted at node u are received at node v at distance r attenuated by a factor $r^{-\alpha/2}h_{u,v}$ for some path-loss exponent $\alpha > 2$, and then further corrupted by additive Gaussian noise. Here $h_{u,v}$ models small-scale fading between the nodes u and v, and is usually assumed to vary in a stationary ergodic fashion across time.

Third, a choice of service requirements has to be made. The simplest such service requirement is *unicast traffic*, in which each message is available at only one source node and requested at only one destination node. When each message is only available at one source node, but the same message may be requested by several destination nodes, we speak of *multicast traffic*. The extreme case, in which each message needs to be transmitted to all the nodes in the network, is termed *broadcast traffic*. Instead of varying the number of destinations for a given message, we can also vary the number of sources that have access to a given message. We think of these sources having access to the same message as caches in the network, replicating these messages. If several sources have access to the same message, but each such message needs to be transmitted to only one destination node, we speak of *caching traffic*.

[2]More commonly, only the first model is called protocol model. The second model is usually referred to as generalized physical model. We use the name protocol model for both of them to highlight that they both make assumptions on the communication protocol used and to contrast them with the more information-theoretic Gaussian fading channel models described in the following.

Fourth, a model for traffic generation is required. The standard assumption for unicast traffic is that each node is source for exactly one other node, and this destination node is chosen uniformly and independently at random from among all the other nodes. Moreover, all these n source-destination pairs generate traffic at equal rate. We refer to this as *random source-destination pairing with uniform rate*. The corresponding maximal achievable per-node rate is called the *throughput capacity* of the wireless network. Another figure of merit for unicast traffic that is often used is the *transport capacity*, which is the maximum achievable rate-distance product, summed over all source-destination pairs. General unicast traffic gives rise to the *unicast capacity region* $\Lambda^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n \times n}$, which characterizes the set of achievable rates for each of the possible n^2 source-destination pairs. For multicast traffic the standard homogeneity assumption is that each node in the network is a source and requires to multicast at uniform rate to the same number of destination nodes chosen uniformly at random. As before, general multicast traffic gives rise to the *multicast capacity region* $\Lambda^{\mathsf{MC}}(n) \subset \mathbb{R}_+^{n \times 2^n}$, which characterizes the set of achievable rates for each of the possible $n \times 2^n$ pairs of source and corresponding group of destinations. Finally, achievable general caching traffic can be described by the *caching capacity region* $\Lambda^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$, which characterizes the set of achievable rates for each of the possible $2^n \times n$ pairs of caches and corresponding destination.

1.2 Prior Work

In this section, we review prior work on scaling laws for wireless networks. Most of the literature on the subject focuses on the homogeneous setting, i.e., random node placement and unicast traffic under random source-destination pairing with uniform rate. The literature pertaining to this homogeneous setting is surveyed in Section 1.2.1. The literature for arbitrary node placement, in which no probabilistic assumptions are made on the node location, is reviewed in Section 1.2.2. Prior work considering more general unicast traffic patterns is discussed in Section 1.2.3. Finally, Section 1.2.4 provides a literature survey for work on service heterogeneity, such as

multicast, broadcast, and caching traffic.

1.2.1 Homogeneous Setting

The scaling approach to analyzing wireless networks was pioneered by Gupta and Kumar in [15]. They show that under random node placement and assuming protocol model 1, the throughput capacity scales like[3] $\Theta\big((n\log(n))^{-1/2}\big)$. For protocol model 2, they prove an upper bound of $O\big(n^{-1/\alpha}\big)$, and a lower bound of $\Omega\big((n\log(n))^{-1/2}\big)$ (see also [14]). Achievability (i.e., the lower bound on the throughput capacity) is shown using a straight-line multi-hop routing scheme. For protocol model 1, a simpler proof of the $\Omega\big((n\log(n))^{-1/2}\big)$ lower bound on the throughput capacity was provided subsequently in [30]. Achievability is shown there by multi-hop routing along a grid structure instead of straight line routing proposed in [15]. Using an argument in [3] relating protocol models 1 and 2, the communication scheme proposed in [30] also applies to protocol model 2. The upper bound for protocol model 2 was later sharpened to $O\big(n^{-1/2}\big)$ in [3]. This leaves only a order $\log^{-1/2}(n)$ gap between the upper and lower bounds for protocol model 2. This gap was closed in [10], where it is shown that, under protocol model 2, the throughput capacity scales like $\Theta\big(n^{-1/2}\big)$. To summarize, under protocol model 1 the throughput capacity scales like $\Theta\big((n\log(n))^{-1/2}\big)$, and under protocol model 2 the throughput capacity scales like $\Theta\big(n^{-1/2}\big)$.

The results mentioned in the last paragraph show that under a protocol model assumption, the maximum achievable per-node rate with random source-destination pairing decays to zero essentially as $\Theta\big(n^{-1/2}\big)$. However, this result was derived by restricting communication schemes to just multi-hop routing (by making the protocol model assumption). While such a restriction is motivated by current technology, it is not clear that multi-hop communication is optimal for large wireless networks. To make claims about the performance of wireless networks under *any* communication scheme, a more information-theoretic approach using a Gaussian channel model (either with or without fading) is necessary.

[3] We use Knuth's asymptotic notation. See Section 2.1 for a formal definition.

Since any communication scheme for the protocol models is also a communication scheme for the Gaussian channel models achieving the same order rate, we obtain from the results mentioned above that under both Gaussian channel models (i.e., with or without fading) throughput capacity is lower bounded by $\Omega\big(n^{-1/2}\big)$. The work on scaling laws under the Gaussian channel models can be grouped into two streams. One stream of work [4, 21, 32, 37, 38, 48, 50, 51] focused on progressively broadening the conditions on the channel model, under which multi-hop communication is indeed order optimal, and hence throughput capacity is also upper bounded by $O\big(n^{-1/2}\big)$. Another stream of work [1, 16, 28, 38, 49] focused on progressively more refined multi-user cooperative schemes, which are shown to significantly out-perform multi-hop communication in certain environments, hence improving the $\Omega\big(n^{-1/2}\big)$ lower bound on the throughput capacity.

For the upper bounds, it was argued in [48] that for the Gaussian channel model with path-loss exponents $\alpha > 6$ (i.e., signal power decays quickly as a function of distance), throughput capacity is upper bounded by $O\big(n^{-1/2}\big)$, and hence multi-hop communication is order optimal in this regime. This result was later extended for the Gaussian fading channel model in [51]. A sharper bound was found subsequently in [21], where it is shown that under both Gaussian channel models (with or without fading), the same upper bound holds for $\alpha > 5$. In [32] it is shown that under a Gaussian channel model, even for $\alpha \in (2, 5]$ the throughput capacity is upper bounded by $O\big(n^{1/\alpha-1/2}\log(n)\big)$. While this does not prove the order optimality of multi-hop communication, it does show that the throughput capacity must decay to zero as $n \to \infty$. The threshold above which multi-hop communication is order optimal was further reduced to $\alpha > 4.5$ in [4], and to $\alpha > 4$ in [50] for both Gaussian channel models. For the Gaussian fading channel model, it is shown in [38] that the throughput capacity is $O\big(n^{-1/2+\varepsilon}\big)$ for $\alpha > 3$ and $O\big(n^{1-\alpha/2+\varepsilon}\big)$ for $\alpha \in (2, 3]$ for any $\varepsilon > 0$. Hence multi-hop is order optimal in the sense of achieving the best scaling exponent for $\alpha > 3$. For the Gaussian channel model without fading, [37] shows that for $\alpha \in (2, 4]$, throughput capacity is upper bounded by $O\big(n^{1/(\alpha+8)-1/2}\log^3(n)\big)$. All these results rely on the cut-set bound to upper bound the sum rate across a cut

by the capacity of a multiple-input multiple-output (MIMO) point-to-point channel in which all the nodes on one side of the cut are allowed to cooperate in sending a message and all nodes on the other side of the cut are allowed to cooperate in receiving this message. They differ, however, in their analysis of this MIMO channel.

For the lower bounds, it was first argued in [16] that there exists a (carefully constructed) node placement such that under either Gaussian channel model (i.e., with or without fading) higher rates than suggested by the results for the protocol models are achievable. This node placement consists of two clusters each containing half the nodes. In the communication scheme proposed in [16], the nodes in the first cluster exchange all their messages among themselves and then jointly encode and transmit them. The first part can be carried out efficiently since the all the nodes are located close to each other. Similarly, the nodes in the second cluster exchange their received observations and then jointly decode them. This procedure effectively transforms the network into a distributed MIMO channel. Similar distributed cooperative schemes were also suggested in [28, 49]. While the results in [16] hold only for a particular node placement, it is shown in [1] that a similar approach can also be used under random node placement. However, since the nodes are now less clustered, setting up the distributed MIMO channel incurs a loss. In [38] it is shown that this loss can be circumvented by using the scheme proposed in [1] multiple times in a hierarchical fashion. More precisely, the problem of setting up the distributed MIMO channel is recognized as being essentially the same as the original communication problem, but at a smaller scale. Using the same scheme recursively, we can thus reduce this scale to a point where the penalty of setting up the initial distributed MIMO channel is negligible. Analyzing this scheme yields that for the Gaussian fading channel model and $\alpha \in (2,3)$, the throughput capacity is lower bounded by $\Omega\big(n^{1-\alpha/2-\varepsilon}\big)$ for any $\varepsilon > 0$. This matches the upper bound derived in the same paper up to ε (see the previous paragraph), and hence establishes that the throughput capacity scales like $\Theta\big(n^{1-\alpha/2\pm\varepsilon}\big)$ in this regime.

To summarize, under the Gaussian channel model with fading, the throughput scales essentially like $\Theta\big(n^{1-\min\{3,\alpha\}/2\pm\varepsilon}\big)$ for any $\alpha > 2$ (with improvements on the $\pm\varepsilon$

Model	Throughput Capacity	
Protocol model 1	$\Theta\big((n\log(n))^{-1/2}\big)$	
Protocol model 2	$\Theta\big(n^{-1/2}\big)$	for $\alpha > 2$
Gaussian with fading	$\Theta\big(n^{-1/2}\big)$	for $\alpha > 4$
	$\Omega\big(n^{-1/2}\big)$	for $\alpha \in (2, 4]$
	$O\big(n^{1/(\alpha+8)-1/2}\log^3(n)\big)$	for $\alpha \in (2, 4]$
Gaussian without fading	$\Theta\big(n^{-1/2}\big)$	for $\alpha > 4$
	$\Omega\big(n^{-1/2}\big)$	for $\alpha \in (3, 4]$
	$O\big(n^{-1/2+\varepsilon}\big)$	for $\alpha \in (3, 4]$ and any $\varepsilon > 0$
	$\Theta\big(n^{1-\alpha/2\pm\varepsilon}\big)$	for $\alpha \in (2, 3)$ and any $\varepsilon > 0$

Table 1.1: Summary of scaling results for the throughput capacity in random wireless networks under various communication/channel models.

possible for $\alpha > 3$). Under the Gaussian channel model without fading the throughput capacity scales like $\Theta\big(n^{-1/2}\big)$ for $\alpha > 4$, and it is lower bounded by $\Omega\big(n^{-1/2}\big)$ and upper bounded by $O\big(n^{1/(\alpha+8)-1/2}\log^3(n)\big)$ for $\alpha \in (2, 4)$. The results reviewed so far for all communication and channel models are listed in Table 1.1.

1.2.2 Location Heterogeneity

Location heterogeneity is usually modeled by allowing arbitrary deterministic node placement with a minimum-separation requirement. For protocol models 1 and 2, such arbitrary node placement can be analyzed by converting the wireless network into an equivalent (wireline) graph, capturing which nodes can communicate with each other, and a set of constraints on the edges that can simultaneously transmit data, capturing the communication constraints of the channel model. This approach is taken in [35], building on results on achievable rates in wireline graphs by [31]. For protocol models 1 and 2, this yields a computable characterization of the capacity scaling for a fairly general set of traffic models.

The situation is more complicated under the Gaussian channel models. Some results on the scaling of the transport capacity (i.e., the maximum achievable rate-distance product summed over all source-destination pairs) under arbitrary node

placement are known. In [48], it is shown that under such node placement and using a Gaussian channel model without fading, the transport capacity is upper bounded by $O(n)$ for $\alpha > 6$. For Gaussian fading channels, the same behavior was shown to hold for $\alpha > 6$ in [51]. Under both Gaussian channel models the same $O(n)$ upper bound on the transport capacity was argued to hold for $\alpha > 5$ in [21], for $\alpha > 4.5$ in [4], and for $\alpha > 4$ in [50]. Matching lower bounds are, however, usually only available under stricter conditions on the node placement. In [51], it is shown that the transport capacity is also lower bounded by $\Omega(n)$ for any $\alpha > 2$ if the node placement is such that for every node at least one other node is within distance $\Theta(1)$. The unicast traffic that achieves this lower bound pairs each node with its nearest neighbor into a source-destination pair, and all these n pairs communicate at equal rate.

1.2.3 Traffic Heterogeneity

As mentioned in the previous section, under protocol models 1 or 2, the wireless network can be transformed into an equivalent wireline graph. This is used in [35] to analyze more general traffic patterns. The authors consider product multicommodity flows, in which each source-destination pair (u, v) wants to communicate at rate $\pi_u \pi_v$, where $\{\pi_u\}$ are arbitrary nonnegative numbers. For such traffic patterns, achievable rates scale like the conductance of the equivalent wireline graph [35]. Another approach is to consider the transport capacity of the wireless network. The transport capacity upper bounds every achievable rate-distance product summed over all source-destination pairs. As such it provides an upper bound on the transport rate for any achievable unicast traffic matrix. In other words, the transport capacity provides a hyperplane that contains the capacity region and origin on the same side. Through a repeated application of this transport capacity bound at different scales, [42, 43] obtained an implicit characterization of the unicast capacity region under a simplified version of protocol model 1. Achievability is shown in [43] using a localized variant of the two-phase Valiant-Brebner routing scheme developed in [47].

For the Gaussian channel models, asymptotic upper bounds on the transport capacity were obtained in [21,48,50,51]. However, as was discussed in the last paragraph,

the transport capacity provides only partial information about the unicast capacity region. Generalized transport capacities, in which the rate between a source-destination pair at distance r is weighted by $f(r)$ for some function f are analyzed in [4]. These generalized transport capacities provide tighter outer bounds on the unicast capacity region.

1.2.4 Service Heterogeneity

So far, we have discussed only unicast traffic. More general service requirements (such as broadcast, multicast, or caching traffic) have recently attracted attention. In [46], broadcasting under (a simplified version of) protocol model 1 is studied. It is shown that under random node placement the maximal per-node rate at which every node can simultaneously broadcast information in the network is upper bounded by $O\big(n^{-1}\big)$. The same problem is analyzed under protocol model 2 in [52], and it is shown that the maximal achievable per-node rate scales like $\Theta\big(n^{-1}\log^{-\alpha/2}(n)\big)$. More general broadcast traffic, in which each node broadcasts data at different (possibly zero) rates, have been studied in [23, 24], where it is shown that general broadcast traffic is achievable if and only if its sum rate scales like $\Theta\big(1\big)$ for protocol model 1 or like $\Theta\big(\log^{-\alpha/2}(n)\big)$ for protocol model 2. In other words, the only relevant quantity when analyzing broadcast traffic is the sum rate. This is because the broadcast requirement induces a uniform *received* traffic pattern, even if the *transmitted* traffic pattern is not (i.e., all nodes are required to receive information at the same rate). An information-theoretic approach to the problem was taken in [40] to analyze broadcast from a single source under random node placement and assuming a Gaussian fading channel model. The maximal achievable broadcast rate for the source is shown to be upper bounded by $O\big(\log\log(n)\big)$ and lower bounded by $\Omega(1)$. Achievability (i.e., the lower bound) is proved using a cooperative multistage scheme. In the first stage, the message is transmitted by the source. In the second stage, nodes that were able to decode the sent message successfully, cooperatively retransmit the message. The scheme continues in the same fashion until all nodes have correctly decoded the message. Similar cooperative schemes for broadcast over Gaussian fading channels

have also been studied in [19, 26].

The analysis of multicast traffic is considerably more difficult. For random node placement and protocol model 1, the maximal uniformly achievable per-node rate for multicast from n^β (with $\beta \in (0,1)$) randomly selected source nodes to the remaining $n^{1-\beta}$ nodes in the network has been shown to scale like $\Theta\big((n^\beta \log(n))^{-1/2}\big)$ in [39]. Under the same assumptions (but with a simplified variant of protocol model 1), it was shown in [33] that when each node wants to multicast at uniform rate to n^β (with $\beta \in (0,1)$) randomly chosen destinations, the maximal achievable per-node rate scales like $\Theta\big((n^{1+\beta} \log(n))^{-1/2}\big)$. This was subsequently generalized to protocol model 2 by [25], where it is shown that under the same traffic and node placement assumptions as in [33], the maximal achievable per node-rate scales like $\Theta\big((n^{1+\beta})^{-1/2}\big)$. Achievability in [25] is shown using the scheme of [10], and the same $\log^{-1/2}(n)$ gap can be observed between the results for protocol models 1 and 2, just as in the unicast case. To the best of our knowledge, the scaling of achievable multicast rates has not been studied from an information-theoretic point of view using either of the Gaussian channel models.

The analysis of caching traffic can be separated into two distinct problems. In the *cache selection* problem, we are given a set of caches and are interested in optimally selecting caches for each destination node and the resulting performance of the network. In the *cache placement* problem, we are interested in optimally placing the caches in order to maximize the performance of the network. Most of the prior work on caching focuses on the second problem and sidesteps the first one by making two assumptions. First, the wireless network is modeled by a (possibly capacitated) graph, and second, each destination node requests the entire message from the closest (with respect to the graph distance) node. For arbitrary graphs, the cache placement problem can then be formulated as a variant of the classical facility location problem (see, e.g., [6, 29] and references therein). In the context of wireless networks, this problem has been studied in [5, 20, 27, 36, 44, 45], with the wireless network modeled as a graph induced by a simplified version of protocol model 1. More precisely, constant factor approximation algorithms for optimal cache placement for one message

under different communication constraints are proposed in [36, 44]. Constant factor approximation algorithms for multiple messages under different memory constraints are proposed in [27,45]. Scaling results for the cache placement problem are presented in [20], which derives asymptotically optimal cache densities assuming random node placement and uniform traffic, and in [5], which analyzes the resulting scaling of achievable rates. As mentioned before, the results on caching traffic surveyed in this paragraph model the wireless network as a graph and assume nearest-neighbor cache selection. Hence they address only the cache placement problem while avoiding the cache selection problem.

Caching in wireless networks has not been directly considered in the information theory literature. However, it can be seen that the cache selection problem is a special case of the problem of communicating correlated sources over a noisy network. Indeed, we can consider that each cache has an identical message to send to the same destination. The more general problem of transmitting correlated sources over noisy networks has received considerable attention. Unlike the situation with point-to-point communication, for network communication problems source-channel separation does not hold in general [8]. Hence, the problem of source and channel coding have to be considered jointly. While for some special cases optimal communication strategies for transmitting correlated sources over a noisy network are known (for example, broadcast from a single source with independent network links [7, 17]), the general problem is unsolved.

1.3 Thesis Outline

This thesis considers the impact of different heterogeneities on achievable rates in a wireless network. Throughout, we assume the Gaussian fading channel model. Our treatment is information theoretic and hence allows claims to be made about the performance of wireless networks under *any* communication scheme consistent with the model.

1.3.1 Location Heterogeneity

As mentioned earlier, the standard homogeneity assumption for the location of nodes is that they are placed independently and uniformly at random on a square of area n. However, in many situations this might not be a good model of reality. A more general assumption is to allow for arbitrary node placement with a constant (independent of n) minimum separation between nodes. We adopt this model to study the effect of location heterogeneity on the scaling of achievable rates. To study this effect in isolation, we keep the homogeneity assumptions for traffic and service requirements, i.e., we assume unicast traffic induced by random source-destination pairing with uniform rate.

Several complications arise due to the introduction of location heterogeneity. As we have seen in Section 1.2.1, in the homogeneous case the optimal communication scheme depends crucially on the path-loss exponent α: For $\alpha \geq 3$, multi-hop communication is order optimal, whereas for $\alpha \in (2, 3]$ hierarchical cooperative communication is order optimal. The first complication is that under arbitrary node placement these schemes might either be clearly suboptimal or not even be implementable at all. As an example, consider the two node placements in Figure 1-1. The left node placement shows infeasibility of hierarchical cooperation under arbitrary node placement. Recall from Section 1.2.1 that the hierarchical cooperation scheme operates by setting up distributed MIMO transmitter and receiver clusters at increasingly bigger scales. In other words, the neighbors of each source help transmitting, and the neighbors of each destination help receiving the message. Consider now the source-destination pair (u, w) in the left node placement in Figure 1-1. These nodes are both isolated, i.e., they have no immediate neighbors. Hence neither the transmitter MIMO cluster nor the receiver MIMO cluster can be effectively constructed. The right node placement shows suboptimality of multi-hop communication. In this example, half of the nodes are placed on the left of the square area, the other half on the right. The gap between these two node clusters is of order $\Theta\big(\sqrt{n}\big)$. Consider the source-destination pair (u, w) in this figure. For a multi-hop communication scheme,

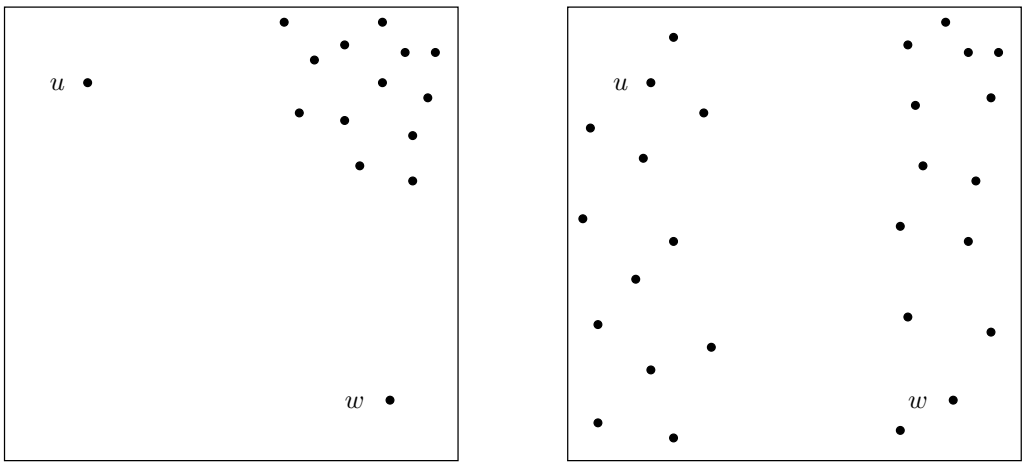

Figure 1-1: Example node placements showing the infeasibility of hierarchical cooperation (left) and the suboptimality of multi-hop communication (right) under arbitrary node placement.

one of the hops will have to cross the gap between the two node clusters. Since this gap is large, this hop will only be able to carry data at low data rates. More precisely, for a hop of size $\Theta\big(\sqrt{n}\big)$, the largest data rate achievable scales like $\Theta\big(n^{-\alpha/2}\big)$. Since under random source-destination pairing a constant fraction of nodes will have to communicate across this gap[4], the maximal uniformly achievable per-node rate under multi-hop communication is at most $O\big(n^{-\alpha/2}\big)$ — much worse than the $\Omega\big(n^{-1/2}\big)$ scaling achievable with multi-hop under random node placement, especially for large α.

To address the infeasibility of the hierarchical cooperative communication scheme under arbitrary node placement, we propose a different hierarchical scheme, which we call *hierarchical relaying*. We show that this scheme achieves the same[5] $n^{1-\alpha/2-o(1)}$ scaling of the per-node rate as hierarchical cooperation, but requires no uniformity in the node placement. In particular, this scheme can be successfully applied to the node placement in the left of Figure 1-1. We also show that for $\alpha \in (2,3]$ the same[6] $O\big(\log^6(n)n^{1-\alpha/2}\big)$ upper bound on the per-node rate (proved in [38] for

[4]For this argument, it is enough if even a single source-destination pair has to cross the gap.

[5]Note that the scaling of the lower bound obtained here is actually slightly better, i.e., $n^{1-\alpha/2-o(1)}$ here compared to $\Omega\big(n^{1-\alpha/2-\varepsilon}\big)$ for arbitrary small, but constant, $\varepsilon > 0$ in [38].

[6]Again, the scaling of the upper bound obtained here is slightly better, i.e., $O\big(\log^6(n)n^{1-\alpha/2}\big)$

the homogeneous case) is also valid under arbitrary node placement. Together, this answers the question of scaling of the throughput capacity under arbitrary node placement for the low path-loss exponent regime $\alpha \in (2, 3]$.

As we argued in the last paragraph, for $\alpha \in (2, 3]$ the node placement has no impact on the scaling performance. The situation is markedly different for large path-loss exponent $\alpha > 3$. We introduce a *regularity parameter*, measuring on a coarse level the uniformity of the node placement. We show how the scaling of throughput capacity depends on this regularity parameter. The proposed order optimal communication scheme smoothly "interpolates" from multi-hop communication (which is order optimal under uniform node placement) to hierarchical relaying (which is order optimal under completely irregular node placement) depending on this regularity parameter. As an example, we show that for the node placement on the right in Figure 1-1, the order optimal communication scheme is hierarchical relaying, achieving a per-node rate of $n^{1-\alpha/2-o(1)}$. This contrasts with the performance of multi-hop communication, which is yields a per-node rate of at most $O\big(n^{-\alpha/2}\big)$.

As mentioned in the introduction, the common approach to deal with heterogeneities consists of identifying the underlying "coarse structure" of the network, capturing the essential parts of the heterogeneity. The coarse structure of the wireless network in this case (for $\alpha > 3$) is a wireline noiseless grid graph where each node in the grid corresponds to a cluster of nodes in the wireless network with cluster size depending on the regularity of the node placement. This coarse structure explicitly captures the amount of cooperation that is required as a function of the regularity of the node placement.

Location heterogeneity is discussed in detail in Chapter 3.

1.3.2 Traffic Heterogeneity

As was discussed in Section 1.1, the standard homogeneity assumption for traffic generation is that each node is source exactly once and wants to transmit data at uniform rate to a destination node chosen uniformly at random from among the other

here compared to $O\big(n^{1-\alpha/2+\varepsilon}\big)$ for any (constant) $\varepsilon > 0$ in [38].

nodes. This kind of traffic pattern has several problematic characteristics. The first one is uniformity of traffic (i.e., all source-destination pairs want to communicate at the same rate). Most traffic patterns observed in large networks (say the Internet) are quite different, in that they have a large number of users generating little traffic, and a small number of users generating a lot of traffic. Traffic variations of this kind are not captured by the homogeneous traffic assumption. The second characteristic of the homogeneous traffic assumption is that, since each node chooses a destination at random, most source-destination pairs will be at a distance of $\Theta\big(\sqrt{n}\big)$. This is again unlike the situation in actual networks where communication patterns are likely to be more localized. The third characteristic of such traffic patterns is that each node is source exactly once and is destination at most a few times. Situations in which we have a server that needs to transmit data to many nodes, or a user downloading data from many other nodes cannot be accommodated under this assumption.

To overcome these limitations, we turn to general unicast traffic. In other words, we are interested in the entire unicast capacity region $\Lambda^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n\times n}$. To study the effect of traffic heterogeneity in isolation, we assume random node placement. As always, we assume a Gaussian fading channel model. While outer bounds on the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ can be derived from results on transport capacity reviewed in Section 1.2.2, these bounds are quite simple in that they only provide one hyperplane containing the capacity region and the origin on one side, and they do not provide a scaling characterization of $\Lambda^{\mathsf{UC}}(n)$. The situation is worse for inner bounds, where except for some special points (as the one resulting from homogeneous traffic) not much is known.

In this thesis, we find inner and outer bounds on the n^2-dimensional unicast capacity region. These bounds behave asymptotically in the same way along at least $n^2 - n$ of the total n^2 dimensions for $\alpha \in (2, 5]$, and for all n^2 dimensions for $\alpha > 5$. Hence they determine the scaling behavior of either most (for $\alpha \in (2, 5]$) or all (for $\alpha > 5$) of the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$. More precisely, we define two sets $\widehat{\Lambda}_1^{\mathsf{UC}}(n), \widehat{\Lambda}_2^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n\times n}$, coinciding along at least $n^2 - n$ dimensions. We show that

for $\alpha \in (2,5]$,

$$n^{-o(1)}\widehat{\Lambda}_1^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_2^{\mathsf{UC}}(n),$$

and for $\alpha > 5$,

$$n^{-o(1)}\widehat{\Lambda}_1^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_1^{\mathsf{UC}}(n).$$

In words, for $\alpha > 5$, if we shrink $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ by a small (in the scaling sense) factor, we obtain an inner bound to the capacity region. If we grow $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ by a small (again in the scaling sense) factor, we obtain an outer bound. Thus $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ scales like $\Lambda^{\mathsf{UC}}(n)$. The same statement is true for $\alpha \in (2,5]$ for $n^2 - n$ out of n^2 dimensions of $\Lambda^{\mathsf{UC}}(n)$. This characterization allows for analysis of the asymptotic behavior of the wireless network under general unicast traffic.

Note that the set $\Lambda^{\mathsf{UC}}(n)$ is large (n^2 dimensional) and could in general be rather difficult to describe. Indeed, descriptions of feasible rates are usually given in terms of cut-set bounds that constrain the sum rate of subsets of nodes. Potentially there are 2^n such subsets, which would result in a very complicated characterization of $\Lambda^{\mathsf{UC}}(n)$. However, we show that the bounds $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ can be described approximately using only $2n$ cuts. More precisely, $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ are polytopes with at most $2n$ faces, each one of them corresponding to some cut-set bound. This shows that, even though the description complexity of $\Lambda^{\mathsf{UC}}(n)$ is likely to be of order $\Theta(2^n)$, the description complexity of its approximation $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ is only of order $\Theta(n)$ — a logarithmic reduction in description complexity!

The coarse structure capturing traffic heterogeneity is a noiseless wireline tree graph. The leaves of this tree correspond to the nodes in the wireless network, intermediate nodes in this tree correspond to various levels of cooperation within the wireless network. This coarse tree structure makes explicit the interaction between traffic demands and the amount of cooperation in the wireless network that is needed to satisfy those demands.

Traffic heterogeneity is discussed in detail in Chapter 4.

1.3.3 Service Heterogeneity

While unicast traffic as discussed in the last two sections describes a broad class of traffic, in several applications multicast is the dominating mode of communication. In multicast traffic each source node wants to transmit its information to a group of destinations. Here we are interested in general multicast traffic, i.e., the multicast capacity region $\Lambda^{\mathsf{MC}}(n) \subset \mathbb{R}_+^{n\times 2^n}$. As in the last section, we assume random node placement. As mentioned in Section 1.2.4, so far the only results available for multicast traffic are under a protocol model assumption and for homogeneous traffic (i.e., each node is source once and wants to communicate to the same number of randomly chosen destinations at uniform rate).

In this thesis, we find inner and outer bounds on the $n \times 2^n$-dimensional multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ under a Gaussian fading channel model. These bounds coincide up to scaling for $n2^n - n$ out of $n2^n$ dimensions for $\alpha \in (2, 5]$ and for all $n2^n$ dimensions for $\alpha > 5$. Hence they determine the scaling behavior of either most (for $\alpha \in (2, 5]$) or all (for $\alpha > 5$) of the multicast capacity region. More precisely, we define two sets $\widehat{\Lambda}_1^{\mathsf{MC}}(n), \widehat{\Lambda}_2^{\mathsf{MC}}(n) \subset \mathbb{R}_+^{n\times 2^n}$, coinciding along at least $n2^n - n$ dimensions. We show that for $\alpha \in (2, 5]$,

$$n^{-o(1)}\widehat{\Lambda}_1^{\mathsf{MC}}(n) \subset \Lambda^{\mathsf{MC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_2^{\mathsf{MC}}(n),$$

and for $\alpha > 5$,

$$n^{-o(1)}\widehat{\Lambda}_1^{\mathsf{MC}}(n) \subset \Lambda^{\mathsf{MC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_1^{\mathsf{MC}}(n),$$

For $\alpha > 5$, this provides a scaling characterization of the entire multicast capacity region, and the same statement holds for $\alpha \in (2, 5]$ along at least $n2^n - n$ dimensions. As before, we show that the approximations $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$ of the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ are described completely by considering only $2n$ out of 2^n possible cut-set bounds. We again make use of the coarse structure of the wireless network developed for traffic heterogeneity (see Section 1.3.2).

In a similar manner, one can analyze the effect of caching traffic, in which a

destination node can obtain the same information from a group of caches. In other words, we are interested in the caching capacity region $\Lambda^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$. We define a set $\widehat{\Lambda}^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$ such that for $\alpha > 6$,

$$n^{-o(1)}\widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n) \subset O(\log^6(n))\widehat{\Lambda}^{\mathsf{CA}}(n),$$

providing a scaling characterization of the complete caching capacity region in the large path-loss regime. Unlike the case for $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$, the caching capacity region cannot be accurately approximated by fewer than 2^n cut-set bounds. However, we show that $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is nevertheless a manageable expression, in that approximate achievability of caching traffic can be evaluated in polynomial time in the description length of caching traffic matrix λ^{CA} (i.e., $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can be checked efficiently even for large networks).

The characterization of the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ provides a complete (approximate) solution to the cache selection problem mentioned in Section 1.2.4 for the high path-loss regime $\alpha > 6$. We hope that this characterization can be used to subsequently optimize over the cache location, which would then also provide an answer to the cache placement problem.

Service heterogeneity is discussed in detail in Chapters 5 and 6.

Chapter 2

Network Model and Notation

In this chapter, we formally define the network and channel models, and give a rigorous definition of the various capacity regions mentioned in Chapter 1.

Section 2.1 introduces some notation used throughout this thesis. Section 2.2 introduces the network and channel models. Section 2.3 formally defines the unicast, multicast, and caching capacity regions.

2.1 Notation and Conventions

We use Knuth's asymptotic notation. For functions $f, g : \mathbb{N} \to \mathbb{R}_+$, we say that

- $f(n) = O(g(n))$ if $\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty$,
- $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$,
- $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$,
- $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

We use the following conventions: K_i for different i, and K, $\widetilde{K}$, ..., denote strictly positive finite constants independent of n. Vectors and matrices are denoted by boldface whenever the vector or matrix structure is of importance. We denote by $(\cdot)^\dagger$ conjugate transpose. To simplify notation, we assume, when necessary, that large real numbers are integers and omit $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ operators. For the same reason,

we also suppress dependence on n within proofs whenever this dependence is clear from the context, and we assume that $n \geq 2$. Throughout, we use $\log(\cdot)$ and $\ln(\cdot)$ for logarithms with respect to base 2 and e, respectively.

2.2 Network Model

Consider the square

$$A(n) \triangleq [0, \sqrt{n}]^2$$

of area n, and let $V(n) \subset A(n)$ be a set of $|V(n)| = n$ nodes on $A(n)$. Each node $v \in V(n)$ represents a wireless device. We make one of the two following assumptions on the node placement. For *random node placement*, we assume that the n nodes $V(n)$ are placed uniformly at random in an independent and identically distributed (i.i.d.) fashion on the area $A(n)$. For *arbitrary node placement*, we make no probabilistic assumptions, but rather assume that $V(n)$ is an arbitrary deterministic node placement such that $r_{u,v} \geq r_{\min}$, where $r_{u,v}$ is the Euclidean distance between u and v, and where $r_{\min} > 0$ is a constant independent of n. Note that, in either case, the node placement is fixed as a function of time. In other words, we assume that the change in location of the nodes in the network is slow enough with respect to the communication delays. We also assume that all node locations are known throughout the entire network.

We use the following channel model. The (sampled) received signal at node v and time t is

$$y_v(t) = \sum_{u \in V(n) \setminus \{v\}} h_{u,v}(t) x_u(t) + z_v(t) \tag{2.1}$$

for all $v \in V(n), t \in \mathbb{N}$, where $\{x_u(t)\}_{u,t}$ is the (sampled) signal sent by the nodes in $V(n)$. Here $\{z_v(t)\}_{v,t}$ are i.i.d. circularly symmetric complex Gaussian random variables with mean 0 and variance 1, and

$$h_{u,v}(t) = r_{u,v}^{-\alpha/2} \exp(\sqrt{-1}\theta_{u,v}(t)), \tag{2.2}$$

for *path-loss exponent* $\alpha > 2$. As a function of the nodes $u, v \in V(n)$, the phases $\{\theta_{u,v}(t)\}_{u,v}$ are assumed to be i.i.d. with uniform distribution on $[0, 2\pi)$. As a function of time t, we either assume that $\{\theta_{u,v}(t)\}_t$ is stationary and ergodic, which is called *fast fading* in the following, or we assume that $\{\theta_{u,v}(t)\}_t$ is constant, which is called *slow fading* in the following. In either case, we assume full channel state information (CSI) is available at all nodes, i.e., each node knows all $\{\theta_{u,v}(t)\}_{u,v}$ at time t. We also impose an average power constraint of 1 on the signal $\{x_u(t)\}_t$ for every node $u \in V(n)$.

While the channel model used is quite simple, it does capture several effects arising in wireless channels. The phase shifts $\{\theta_{u,v}(t)\}_{u,v}$ model the effect of small-scale movements of the nodes (on the order of the wavelength). The i.i.d. assumption of the phase shifts is justified by the large (again, relative to the wavelength) separation of the nodes (but see the comments on the validity of the model for very large n and $\alpha \in (2, 3)$ below). The $r_{u,v}^{-\alpha/2}$ term models power decay over larger scales, and is assumed not to be affected by the small-scale movement. Since the network is assumed to be static, the $r_{u,v}^{-\alpha/2}$ terms do not vary with time.

The full CSI assumption made above is quite strong, and is worth commenting on. First, we make the full CSI assumption in all the converse results in this thesis. This implies that all the converses also hold under weaker assumptions on the CSI, and hence are valid as well under a wide variety of more realistic assumptions on the availability of side information. Second, all achievability results presented in this thesis can be shown to hold under weaker assumptions on the availability of CSI. In all cases, a 2-bit quantization of the channel state $\{\theta_{u,v}(t)\}_{u,v}$ available at all nodes in $V(n)$ at time t is sufficient to obtain the same scaling behavior. Moreover, for random node placement and $\alpha \in (2, 3]$, causal quantized *receiver only* CSI is sufficient. And for random node placement and $\alpha \geq 3$ no CSI is needed. We comment on the necessity of CSI in more detail following the proofs of the scaling results in subsequent chapters.

We should also point out that recent results [11] suggest that, under certain assumptions on the location of scattering elements, for $\alpha \in (2, 3)$ and very large values of n, the channel model used here (in particular, the i.i.d. assumption on the phases

$\{\theta_{u,v}(t)\}_{u,v}$ as a function of the nodes $u, v \in V(n)$) might yield results that are too optimistic. However, the authors show in [12] that, under different assumptions on the scatterers, the channel model used here is still valid also for $\alpha \in (2,3)$ and very large values of n. This indicates that the issue of proper channel modelling in the low path-loss regime for very large networks is somewhat delicate and requires further investigation.

2.3 Capacity Regions

A *traffic matrix* $\lambda \in \mathbb{R}_+^{2^n \times 2^n}$ associates with each pair of subsets $U, W \subset V(n)$ of nodes the number $\lambda_{U,W}$. This $\lambda_{U,W}$ is to be understood as the *rate* at which the nodes in W request a common message available at the set of caches U. We are interested in the set of traffic matrices that the wireless network can support. The collection of all such supportable traffic matrices will be called the *capacity region* $\Lambda(n) \subset \mathbb{R}^{2^n \times 2^n}$ of the wireless network.

We now make the definition of $\Lambda(n)$ formal. Fix a traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times 2^n}$ and a *blocklength* $T \in \mathbb{N}$. Let the *message* $m_{U,W}^{(T)}$ be uniformly distributed on

$$\left\{1, \ldots, 2^{T\lambda_{U,W}}\right\}.$$

We assume that the random variables $\{m_{U,W}^{(T)}\}_{U,W \subset V(n)}$ are independent. Note that $m_{U,W}^{(T)}$ is requested by all destination nodes $w \in W$ and is available at all nodes $u \in U$. Hence node u has access to all messages $m_{U,W}^{(T)}$ such that $u \in U$, i.e.,

$$\{m_{U,W}^{(T)}\}_{U,W \subset V(n): u \in U}.$$

The *message set* at node $u \in V(n)$ is then defined as the set of all possible values of these message available at u:

$$M_u^{(T)} \triangleq \bigotimes_{U \subset V(n): u \in U} \bigotimes_{W \subset V(n)} \left\{1, \ldots, 2^{T\lambda_{U,W}}\right\} \quad \forall u \in V(n).$$

An *encoder of blocklength* T is a collection of functions

$$x_u^{(T)}(t) : M_u^{(T)} \times [0, 2\pi)^{tn(n-1)} \times \mathbb{C}^{t-1} \to \mathbb{C} \quad \forall t \in \{1, \dots T\}, u \in V(n),$$

mapping the messages $\{m_{U,W}^{(T)}\}$ available at u (i.e., satisfying $u \in U$), the channel states $\{\{\theta_{u,v}(s)\}_{u,v\in V(n)}\}_{s=1}^{t}$ up to time t, and the received signals $\{y_u(s)\}_{s=1}^{t-1}$ at node u up to time $t-1$ into a channel input $x_u^{(T)}(t)$ at time t. We impose that the encoder satisfies the power constraint

$$\sum_{\{m_{U,W}^{(T)}\}\in M_u^{(T)}} \sum_{t=1}^{T} \frac{1}{T|M_u^{(T)}|} \mathbb{E}\Big(|x_u^{(T)}(t)\big(\{m_{U,W}^{(T)}\}, \{\{\theta_{u,v}(s)\}_{u,v\in V(n)}\}_{s=1}^{t}, \{y_u(s)\}_{s=1}^{t-1}\big)|^2\Big)$$
$$\leq 1 \quad \forall u \in V(n),$$

with expectation with respect to $\{m_{U,W}\}$, $\{y_u(s)\}_{s=1}^{t-1}$ and $\{\{\theta_{u,v}(s)\}_{u,v\in V(n)}\}_{s=1}^{t}$. A *decoder of blocklength* T is a collection of functions

$$\varphi_{U,W,w}^{(T)} : \mathbb{C}^T \times [0, 2\pi)^{Tn(n-1)} \times M_w^{(T)} \to \big\{1, \dots, 2^{T\lambda_{U,W}}\big\} \quad \forall U, W \subset V(n) : w \in W,$$

mapping the received signal $\{y_w(t)\}_{t=1}^{T}$ at node w up to time T, the channel states $\{\{\theta_{u,v}(t)\}_{u,v\in V(n)}\}_{t=1}^{T}$ up to time T, and the messages $\{m_{\widetilde{U},\widetilde{W}}^{(T)}\}$ available at node w (i.e., satisfying $w \in \widetilde{U}$) into an estimate $\hat{m}_{U,W}^{(T)}$ of the message $m_{U,W}^{(T)}$. Together, an encoder and a decoder of blocklength T form a *coding scheme of blocklength* T. The *probability of error* or such a coding scheme is defined as

$$P_e^{(T)} \triangleq$$
$$\max_{U,W\subset V(n)} \max_{w\in W} \mathbb{P}\Big(\varphi_{U,W,w}^{(T)}\big(\{y_w(t)\}_{t=1}^{T}, \{\{\theta_{u,v}(t)\}_{u,v}\}_{t=1}^{T}, \{m_{\widetilde{U},\widetilde{W}}^{(T)}\}_{\widetilde{U},\widetilde{W}:w\in\widetilde{U}}\big) \neq m_{U,W}^{(T)}\Big).$$

In words, $P_e^{(T)}$ is the average probability of error of incorrect decoding of the message $m_{U,W}^{(T)}$ maximized over all possible caches U and destination groups W.

A traffic matrix $\lambda \in \mathbb{R}_+^{2^n\times 2^n}$ is said to be *achievable*, if there exists a sequence of

coding schemes of blocklengths $T \in \mathbb{N}$ such that

$$\lim_{T\to\infty} P_e^{(T)} = 0.$$

Finally, the *capacity region* $\Lambda(n) \subset \mathbb{R}_+^{2^n \times 2^n}$ is the closure of the set of all achievable traffic matrices.

A few examples will illustrate the utility of defining the notion of the capacity region $\Lambda(n)$ in such generality. These examples introduce important special cases that will be analyzed throughout this thesis.

Example 2.1. (*Unicast*)

A *unicast traffic matrix* $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n\times n}$ associates with each node pair $u, w \in V(n)$ a number $\lambda_{u,w}^{\mathsf{UC}}$. This number is the rate at which the source node u wants to transmit information to the destination node w. Note that we allow the same node u to be source for multiple destinations and the same node w to be destination for multiple sources. In such situations, the multiple messages at u are assumed to be independent (and similarly for the messages from multiple sources at w).

For a specific example, assume $n = 4$, and label the nodes as $\{u_i\}_{i=1}^4 = V(n)$. Assume further that node u_1 needs to transmit a message m_{u_1,u_2} to node u_2 at rate 1 bit per channel use and an independent message m_{u_1,u_3} to node u_3 at rate 2 bits per channel use. Node u_2 needs to transmit a message m_{u_2,u_3} to node u_3 at rate 4 bits per channel use. All the messages $m_{u_1,u_2}, m_{u_1,u_3}, m_{u_2,u_3}$ are independent. This traffic pattern can be described by a unicast traffic matrix $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{4\times 4}$ with $\lambda_{u_1,u_2}^{\mathsf{UC}} = 1$, $\lambda_{u_1,u_3}^{\mathsf{UC}} = 2$, $\lambda_{u_2,u_3}^{\mathsf{UC}} = 4$, and $\lambda_{u,v}^{\mathsf{UC}} = 0$ otherwise. Note that in this example node u_1 is source for two (independent) messages, and node u_3 is destination for two (again independent) messages. Node u_4 in this example is neither source nor destination for any message, and can be understood as a helper node.

Now, for each unicast traffic matrix $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n\times n}$, we can construct a traffic matrix

$\lambda \in \mathbb{R}_+^{2^n \times 2^n}$ as

$$\lambda_{U,W} = \begin{cases} \lambda_{u,w}^{\mathsf{UC}} & \text{if } U = \{u\}, W = \{w\} \text{ for some } u, w \in V(n), \\ 0 & \text{else.} \end{cases}$$

A unicast traffic matrix λ^{UC} is *achievable* if the corresponding traffic matrix λ is. The *unicast capacity region* $\Lambda^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n \times n}$ is defined as the closure of the set of achievable unicast traffic matrices. Note that the unicast capacity region is the subset of the capacity region arising from intersecting $\Lambda(n)$ with the subspace corresponding to (U, W) pairs of the form $(\{u\}, \{w\})$ for some $u, w \in V(n)$.

The notion of unicast traffic defined in the last paragraph is very general. Two special cases of unicast traffic matrices are, however, worth mentioning.

A unicast traffic matrix λ^{UC} is called a *permutation traffic matrix* if for every node $u \in V(n)$ there is exactly one $w \in V(n) \setminus \{u\}$ such that $\lambda_{u,w}^{\mathsf{UC}} > 0$ and exactly one $\tilde{w} \in V(n) \setminus \{u\}$ such that $\lambda_{\tilde{w},u}^{\mathsf{UC}} > 0$. In words, for a permutation unicast traffic matrix, every node is source and destination exactly once. A permutation unicast traffic matrix is said to have *uniform rate* if for all $u, w \in V(n)$ we have $\lambda_{u,w}^{\mathsf{UC}} = \{0, 1\{$ (i.e, each of the n source-destination pairs wants to transmit messages at rate 1). For a permutation traffic matrix λ^{UC} with uniform rate, we define the *throughput capacity* $\rho^*(n)$ as the largest value of $\rho(n)$ such that $\rho(n)\lambda^{\mathsf{UC}}$ is achievable. In other words $\rho^*(n)$ is the largest uniformly achievable per-node rate. For ease of notation, we will often just refer to the throughput capacity $\rho^*(n)$ for a permutation traffic matrix λ^{UC} without explicit mentioning of the uniform rate requirement.

A unicast traffic matrix λ^{UC} is called a *random source-destination pairing with uniform rate* if it results from picking for each node $u \in V(n)$ one other node w independently and uniformly at random from $V(n) \setminus \{u\}$ and setting $\lambda_{u,w}^{\mathsf{UC}} = 1$. Random source-destination pairings with uniform rate are closely related to permutation traffic with uniform rate for which all source-destination pairs are at a distance $\Theta(\sqrt{n})$, and for scaling purposes the two are equivalent. $\Diamond$

Example 2.2. (*Multicast*)

A *multicast traffic matrix* $\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n \times 2^n}$ associates with each node $u \in V(n)$ and subset $W \subset V(n)$ a number $\lambda^{\mathsf{MC}}_{u,W}$. This number is the rate at which the source node u wants to multicast (identical) information to *all* the destination nodes $w \in W$. Note that we do not impose that a source node u multicasts information only to one group of destinations W. In fact, for every $u \in V(n)$ there could be two (or more) subsets $W, \widetilde{W} \subset V(n)$ with $W \neq \widetilde{W}$ such that $\lambda^{\mathsf{MC}}_{u,W} > 0$ and $\lambda^{\mathsf{MC}}_{u,\widetilde{W}} > 0$. In such a situation, the messages for the two groups of destinations are assumed to be independent. Similarly, two nodes could want to multicast (independent) messages to the same set of destination nodes.

For a specific example, assume again $n = 4$, and label the nodes as $\{u_i\}_{i=1}^4 = V(n)$. Assume that node u_1 needs to transmit the same message $m_{u_1,\{u_2,u_3,u_4\}}$ to all nodes u_1, u_2, u_3 at a rate of 1 bit per channel use and an independent message $m_{u_1,\{u_2\}}$ to only node 2 at rate 2 bits per channel use. Node 2 needs to transmit a message $m_{u_2,\{u_1,u_3\}}$ to both u_1, u_3 at rate 4 bits per channel use. All the messages $m_{u_1,\{u_2,u_3,u_4\}}, m_{u_1,\{u_2\}}, m_{u_2,\{u_1,u_3\}}$ are independent. This traffic pattern can be described by a multicast traffic matrix $\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{4 \times 16}$ with $\lambda^{\mathsf{MC}}_{u_1,\{u_2,u_3,u_4\}} = 1$, $\lambda^{\mathsf{MC}}_{u_1,\{u_2\}} = 2$, $\lambda^{\mathsf{MC}}_{u_2,\{u_1,u_3\}} = 4$, and $\lambda^{\mathsf{MC}}_{u,W} = 0$ otherwise. Note that in this example node u_1 is source for two (independent) multicast messages, and node u_2 and u_3 are destinations for more than one message. The message $m_{u_1,\{u_2,u_3,u_4\}}$ is destined for the all the nodes in the network, and can hence be understood as a broadcast message. The message $m_{u_1,\{u_2\}}$ is only destined for one node, and can hence be understood as a private message.

For each multicast traffic matrix $\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n \times 2^n}$, we can construct a traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times 2^n}$ as

$$\lambda_{U,W} = \begin{cases} \lambda^{\mathsf{MC}}_{u,W} & \text{if } U = \{u\} \text{ for some } u \in V(n), \\ 0 & \text{else.} \end{cases}$$

A multicast traffic matrix λ^{MC} is *achievable* if the corresponding traffic matrix λ is. The *multicast capacity region* $\Lambda^{\mathsf{MC}}(n) \subset \mathbb{R}_+^{n \times 2^n}$ is defined as the closure of the set of achievable unicast traffic matrices. As before, the multicast capacity region is

the subset of the capacity region arising from intersecting $\Lambda(n)$ with the subspace corresponding to (U, W) pairs of the form $(\{u\}, W)$ for some $u \in V(n)$.

We note that this definition of multicast is very general. $\Diamond$

Example 2.3. (*Caching*)

A *caching traffic matrix* $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ associates with each node $w \in V(n)$ and subset $U \subset V(n)$ a number $\lambda^{\mathsf{MC}}_{U,w}$. This number is the rate at which the destination node w requests information that is available at *all* the caches $u \in U$. Note that we do not impose that a destination node w requests information from only one group of caches U. In fact, for every $w \in V(n)$ there could be two (or more) subsets $U, \widetilde{U} \subset V(n)$ with $U \neq \widetilde{U}$ such that $\lambda^{\mathsf{CA}}_{U,w} > 0$ and $\lambda^{\mathsf{CA}}_{\widetilde{U},w} > 0$. In such a situation, the messages for the two groups of caches are assumed to be independent. Similarly, the same set of caches can hold (independent) messages for more than one different destination nodes. For example, a situation where parts of a message requested by a destination node w is available at caches U and a different part is available at caches $\widetilde{U}$ could be modeled as two messages (one corresponding to each part) available at U and $\widetilde{U}$, respectively.

For a specific example consider again $\{u_i\}_{i=1}^4 = V(n)$ with $n = 4$. Assume that u_1 requests a message $m_{\{u_3,u_4\},u_1}$ available at the caches u_3, and u_4 at rate 1 bit per channel use and an independent message $m_{\{u_3\},u_1}$ available only at u_3 at a rate of 2 bits per channel use. Node u_2 requests a message $m_{\{u_3,u_4\},u_2}$ available at the caches u_3 and u_4 at a rate of 4 bits per channel use. The messages $m_{\{u_3,u_4\},u_1}$, $m_{\{u_3\},u_1}$, and $m_{\{u_3,u_4\},u_2}$ are assumed to be independent. This traffic pattern can be described by a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{16\times 4}$ with $\lambda^{\mathsf{CA}}_{\{u_3,u_4\},u_1} = 1$, $\lambda^{\mathsf{CA}}_{\{u_3\},u_1} = 2$, $\lambda^{\mathsf{CA}}_{\{u_3,u_4\},u_2} = 4$, and $\lambda^{\mathsf{CA}}_{U,w} = 0$ otherwise. Note that in this example node u_1 is destination for two (independent) caching messages, and node u_3 and u_4 serve as caches for more than one message (but these messages are assumed independent).

For each caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$, we can construct a traffic matrix

$\lambda \in \mathbb{R}_+^{2^n \times 2^n}$ as

$$\lambda_{U,W} = \begin{cases} \lambda_{U,w}^{\mathsf{CA}} & \text{if } W = \{w\} \text{ for some } w \in V(n), \\ 0 & \text{else.} \end{cases}$$

A caching traffic matrix λ^{CA} is *achievable* if the corresponding traffic matrix λ is. The *caching capacity region* $\Lambda^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$ is defined as the closure of the set of achievable caching traffic matrices. As before, the caching capacity region is the subset of the capacity region arising from intersecting $\Lambda(n)$ with the subspace corresponding to (U, W) pairs of the form $(U, \{w\})$ for some $w \in V(n)$.

This definition of caching is completely general in terms of the number and location of caches and their destinations as well as the amounts of traffic between them. Moreover, by the definition of achievability, we do not impose that for a particular (U, w) pair one cache $u \in U$ transmits the entire requested message to the destination node w. Rather, we allow all caches to participate in the transmission of the message. Thus, this definition of caching is also general in terms of cache selection. $\lozenge$

We note that the definition of the capacity region (and hence also the ones for unicast, multicast, and caching) contain several trivial dimensions. These are the dimensions corresponding to (U, W) pairs such that either $W \subset U$ with $W \neq \emptyset$, or $U = \emptyset$, or $W = \emptyset$. The first such case can arise in unicast, multicast, and caching and corresponds to $w = u$, $W = \{u\}$, and $w \in U$, respectively. The second case arises only in caching. The third case arises only in multicast. We now analyze these three trivial cases in more detail.

Consider an entry $\lambda_{U,W}$ of the traffic matrix λ such that $W \subset U$. Note that the decoder $\varphi_{U,W,w}^{(T)}$ at node $w \in W$ has access to the messages $\{m_{\widetilde{U},\widetilde{W}}^{(T)}\}_{\widetilde{U},\widetilde{W}:w\in\widetilde{U}}$. In particular, since $W \subset U$ and hence $w \in U$, it has access to $m_{U,W}^{(T)}$, and can therefore easily decode this message by simply setting

$$\varphi_{U,W,w}^{(T)}\big(\{y_w(t)\}_{t=1}^T, \{\{\theta_{u,v}(t)\}_{u,v\in V(n)}\}_{t=1}^T, \{m_{\widetilde{U},\widetilde{W}}^{(T)}\}_{\widetilde{U},\widetilde{W}:w\in\widetilde{U}}\big) = m_{U,W}^{(T)}.$$

Since this is true for every $w \in W$, we can choose $\lambda_{U,W}$ arbitrarily large and still guarantee successful decoding. Hence the capacity region $\Lambda(n)$ is unbounded along dimension (U, W) for $W \subset U$.

Consider now an entry $\lambda_{U,W}$ of a traffic matrix λ such that $U = \emptyset$. Then $m_{U,W}^{(T)} \notin M_u^{(T)}$ for any $u \in V(n)$, and therefore no encoder $x_u^{(T)}(t)$ has access to $m_{U,W}^{(T)}$. Hence the received signal at any decoder $\varphi_{U,W,w}^{(T)}$ for $w \in W$ is independent of $m_{U,W}^{(T)}$ and the resulting probability of error will be approaching one as $T \to \infty$ unless $\lambda_{U,W} = 0$. Thus the capacity region $\Lambda(n)$ is zero along dimension (U, W) for $U = \emptyset$.

Finally, consider an entry $\lambda_{U,W}$ of a traffic matrix λ such that $W = \emptyset$. Then there exists no decoder $\varphi_{U,W,w}^{(T)}$ such that $w \in W$, and therefore we can choose $\lambda_{U,W}$ arbitrarily large without affecting the probability of error. Hence the capacity region $\Lambda(n)$ is unbounded along dimension (U, W) for $W = \emptyset$.

While the capacity region $\Lambda(n)$ and all its special cases have certain dimensions that are trivial, these are only very few. In particular, for the $n \times n$ dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ only n dimensions are trivial, for the $n \times 2^n$ dimensional multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ only $2n$ dimensions are trivial, for the $2^n \times n$ dimensional caching capacity region $\Lambda^{\mathsf{CA}}(n)$ only $n(2^{n-1} + 1)$ dimensions are trivial. In other words, the nontrivial number of dimensions of the unicast, multicast, and caching capacity regions are $n(n-1)$, $n(2^n - 2)$, and $n(2^{n-1} - 1)$, respectively. Thus the number of trivial dimensions is negligible, and including them in the definition allows to simplify notation considerably.

Note that the capacity region $\Lambda(n)$ is (in most cases, see below) a random variable with probabilistic structure determined by the assumptions on the node placement and the fading model. More precisely, for slow fading (in which the channel gains are random across nodes, but constant across time), $\Lambda(n)$ is a function of the realization of those channel gains. In contrast, for the fast fading case (in which the channel gains are ergodic across time), the coding scheme can average out any short time fluctuations in the channel gains, and hence $\Lambda(n)$ depends only on the expected behavior of the channel gains and not on their realization. The capacity region $\Lambda(n)$ is always a function of the node placement. However, this only introduces

randomness into the behavior of $\Lambda(n)$ if the node placement is itself random (as opposed to arbitrary deterministic node placement). Finally, $\Lambda(n)$ never depends on the realization of the noise process, as this process is always assumed to be ergodic.

Chapter 3

Location Heterogeneity

In this chapter, we analyze the impact of location heterogeneity on the performance of a wireless network. To this end, we consider wireless networks with arbitrary (i.e., deterministic) node placement (with minimum-separation constraint). As a measure of performance, we use the throughput capacity $\rho^*(n)$ under permutation traffic (i.e. each node is source and destination for exactly one pair, and there are n such source-destination pairs with uniform traffic demand). Before we proceed, recall that under random node placement the throughput capacity scales like $\rho^*(n) = n^{1-\min 3,\alpha/2\pm o(1)}$, and that for small path-loss exponents $\alpha \in (2,3]$ cooperative communication is order optimal and for large path-loss exponents $\alpha > 3$ multi-hop communication is order optimal.

The impact of this arbitrary node placement depends crucially on the path-loss exponent α. For small path-loss exponents $\alpha \in (2,3]$, we show that for random source-destination pairing, the throughput capacity is upper bounded as $\rho^*(n) = O(\log^6(n)n^{1-\alpha/2})$. We then present a novel cooperative communication scheme that achieves for any node placement and path-loss exponent $\alpha > 2$ a per-node rate of $n^{1-\alpha/2-o(1)}$. Thus, our cooperative communication scheme is essentially order optimal for any such arbitrary network with $\alpha \in (2,3]$. In other words, in the small path-loss regime, the scaling of $\rho^*(n)$ is the same irrespective of the regularity of the node placement.

The situation is, however, quite different for large path-loss exponents $\alpha > 3$.

We show that in this regime the scaling of $\rho^*(n)$ depends crucially on the regularity of the node placement, and multi-hop communication may not be order optimal for any value of α. In fact, for less regular networks we need more complicated cooperative communication schemes to achieve optimal network performance. Towards that end, we present a family of communication schemes that smoothly "interpolate" between cooperative communication and multi-hop communication, and in which nodes communicate at scales that vary smoothly from local to global. The amount of "interpolation" between the cooperative and multi-hop schemes depends on the level of regularity of the underlying node placement. We establish the optimality of this family of schemes for all $\alpha > 3$ under adversarial node placement with regularity constraint.

The remainder of this chapter is organized as follows. Section 3.1 provides formal statements of our results. Sections 3.2 and 3.3 describe our new cooperative communication scheme (for the $\alpha \in (2, 3]$ regime) and "interpolation" scheme (for the $\alpha > 3$ regime) for arbitrary wireless networks. Sections 3.4 through 3.9 contain proofs. Finally, Sections 3.10 and 3.11 contain discussions and concluding remarks.

3.1 Main Results

This section presents the formal statement of our results. In Section 3.1.1, we consider low path-loss exponents, i.e., $\alpha \in (2, 3]$. We present a cooperative communication scheme for arbitrary node placement and for either fast or slow fading. We show that this communication scheme is order optimal for all node placements when $\alpha \in (2, 3]$. In Section 3.1.2, we consider high path-loss exponents, i.e., $\alpha > 3$. We present a communication scheme that "interpolates" between the cooperative and the multi-hop communication schemes, depending on the regularity of the node placement. We show that this communication scheme is order optimal under adversarial node placement with regularity constraint when $\alpha > 3$.

3.1.1 Low Path-Loss Regime $\alpha \in (2, 3]$

The first result proposes a novel cooperative communication scheme, called *hierarchical relaying* in the following, and bounds the per-node rate $\rho^{\mathsf{HR}}(n)$ that it achieves. This provides a lower bound to $\rho^*(n)$, the largest achievable per-node rate. The hierarchical relaying scheme enables cooperative communication on the scale of the network size. In the random node placement case, this cooperation could be enabled in a cluster around the source node (cooperatively transmitting) and in a cluster around its destination node (cooperatively receiving). With arbitrary node placement, such an approach no longer works, as both the source as well as the destination nodes may be isolated. The hierarchical relaying scheme circumvents this issue by relaying data between each source-destination pair over a densely populated region in the network. A detailed description of this scheme is provided in Section 3.2, the proof of Theorem 3.1 is contained in Section 3.5.

Theorem 3.1. *Under fast fading, for any $\alpha > 2$, $r_{\min} \in (0, 1)$, and $\delta \in (0, 1/2)$, there exists*

$$b_1(n) \geq n^{-O\left(\log^{\delta-1/2}(n)\right)}$$

such that for any n, node placement $V(n)$ with minimum separation $r_{\min}$, and permutation traffic matrix $\lambda^{\mathsf{UC}}(n)$, we have

$$\rho^*(n) \geq \rho^{\mathsf{HR}}(n) \geq b_1(n) n^{1-\alpha/2}.$$

The same conclusion holds for slow fading with probability at least

$$1 - \exp\Big(- 2^{\Omega\left(\log^{1/2+\delta}(n)\right)}\Big) = 1 - o(1)$$

as $n \to \infty$.

Theorem 3.1 shows that the per-node rate $\rho^{\mathsf{HR}}(n)$ achievable by the hierarchical relaying scheme is at least $n^{1-\alpha/2-\beta(n)}$, where the "loss" term $\beta(n)$ converges to zero as $n \to \infty$ at a rate arbitrarily close to $O\big(\log^{-1/2}(n)\big)$ (by choosing δ small). The

performance of the hierarchical relaying scheme can intuitively be understood as follows. As mentioned before, the scheme achieves cooperation on a global scale. This leads to a multi-antenna gain of order n. On the other hand, communication is over a distance of order $n^{1/2}$, leading to a power loss of order $n^{-\alpha/2}$. Combining these two factors results in a per-node rate of $n^{1-\alpha/2}$.

We note that Theorem 3.1 remains valid under somewhat weaker conditions than having minimum separation $r_{\min} \in (0,1)$. Specifically, we show that the result of Özgür et al. [38] can be recovered through Theorem 3.1 as the random node placement satisfies these weaker conditions. We discuss this in more detail in Section 3.10.4.

The next theorem establishes optimality of the hierarchical relaying scheme in the range of $\alpha \in (2,3]$ for arbitrary node placement. The proof of the theorem is presented in Section 3.6.

Theorem 3.2. *Under either fast or slow fading, for any $\alpha \in (2,3]$, $r_{\min} \in (0,1)$, there exists $b_2(n) = O\big(\log^6(n)\big)$ such that for any n, node placement $V(n)$ with minimum separation $r_{\min}$, and for $\lambda^{\mathsf{UC}}(n)$ chosen uniformly at random from the set of all permutation traffic matrices, we have*

$$\rho^*(n) \leq b_2(n) n^{1-\alpha/2}$$

with probability $1 - o(1)$ as $n \to \infty$.

Note that Theorem 3.2 holds only with probability $1 - o(1)$ for different reasons for the slow and fast fading case. For fast fading, this is due to the randomness in the selection of the permutation traffic matrix. In other words, for fast fading, with high probability we select a traffic matrix for which the theorem holds. For the slow fading case, there is additional randomness due to the fading realization. Here, with high probability we select a traffic matrix and we experience a fading for which the theorem holds.

Comparing Theorems 3.1 and 3.2, we see that for $\alpha \in (2,3]$ the proposed hierar-

chical relaying scheme is order optimal, in the sense that

$$\lim_{n\to\infty} \frac{\log(\rho^{\mathsf{HR}}(n))}{\log(n)} = \lim_{n\to\infty} \frac{\log(\rho^{*}(n))}{\log(n)} = 1 - \alpha/2.$$

Moreover, the rate it achieves is the same order as is achievable in the case of randomly placed nodes. Hence in the low path-loss regime $\alpha \in (2, 3]$, the heterogeneity caused by the arbitrary node placement has no effect on achievable communication rates.

3.1.2 High Path-Loss Regime $\alpha > 3$

We now turn to the high path-loss regime $\alpha > 3$. In the case of *randomly* placed nodes, multi-hop communication achieves a per-node rate of $\rho^{\mathsf{MH}}(n) = \Omega(n^{-1/2})$ with probability $1 - o(1)$ and is order optimal for $\alpha > 3$. For *arbitrarily* placed nodes, the situation is quite different as Theorem 3.3 shows. The proof of Theorem 3.3 is contained in Section 3.7.

Theorem 3.3. *Under either fast or slow fading, for any $\alpha > 3$, for any n, there exists a node placement $V(n)$ with minimum separation $1/2$ such that for $\lambda^{\mathsf{UC}}(n)$ chosen uniformly at random from the set of all permutation traffic matrices, we have*

$$\rho^{*}(n) \leq 2^{2+5\alpha} n^{1-\alpha/2},$$
$$\rho^{\mathsf{MH}}(n) \leq 4^{\alpha} n^{-\alpha/2},$$

as $n \to \infty$ with probability $1 - o(1)$.

Comparing Theorem 3.3 with Theorem 3.1 shows that under adversarial node placement with minimum-separation constraint the hierarchical relaying scheme is order optimal even when $\alpha > 3$. Moreover, the theorems show that there exist node placements satisfying a minimum separation constraint for which hierarchical relaying achieves a rate of at least a factor of order n higher than multi-hop communication for any $\alpha > 3$. In other words, for those node placements cooperative communication is necessary for order optimality also for any $\alpha > 3$, in stark contrast to the situation

with random node placement, where multi-hop communication is order optimal for all $\alpha > 3$.

Theorem 3.3 suggests that it is the level of regularity of the node placement that decides what scheme to choose for path-loss exponent $\alpha > 3$. So far, we have seen two extreme cases: For random node placement, resulting in very regular node placements with high probability, only local cooperation is necessary and multi-hop is an order-optimal communication scheme. For adversarial arbitrary node placement, resulting in a very irregular node placement, global cooperation is necessary and hierarchical relaying is an order-optimal communication scheme. We now make this notion of regularity precise and introduce a *cooperative multi-hop* communication scheme that "interpolates" between multi-hop communication and hierarchical relaying depending on the regularity of the node placement.

Before we state the result, we need to introduce some notation. Consider again a node placement $V(n) \subset A(n)$ with minimum separation $r_{\min} \in (0, 1)$. Divide $A(n)$ into squares of sidelength $d(n) \leq \sqrt{n}$, and fix a constant $\mu \in (0, 1]$. We say that $V(n)$ *is μ-regular at resolution $d(n)$* if every such square contains at least $\mu d^2(n)$ nodes. Note that every node placement is trivially 1-regular at resolution $\sqrt{n}$; a random node placement can be shown to be μ-regular at resolution $\log(n)$ with probability $1 - o(1)$ as $n \to \infty$ for any $\mu < 1$; and nodes that are placed on each point in the integer lattice inside $A(n)$ are 1-regular at resolution 1.

The cooperative multi-hop scheme enables cooperative communication on the scale of regularity $d(n)$. Neighboring squares of sidelength $d(n)$ cooperatively communicate with each other. To transmit between a source and its destination, we use multi-hop communication over those squares. In other words, we use cooperative communication at small scale $d(n)$, and multi-hop communication at large scale $\sqrt{n}$. For regular node placements, i.e., $d(n) = 1$, the cooperative multi-hop scheme becomes the classical multi-hop scheme. For very irregular node placement, i.e., $d(n) = n^{1/2}$, the cooperative multi-hop scheme becomes the hierarchical relaying scheme discussed in the last section.

The next theorem provides a lower bound on the per-node rate $\rho^{\mathsf{CMH}}(n)$ achievable

with the cooperative multi-hop scheme. The proof of the theorem can be found in Section 3.8.

Theorem 3.4. *Under fast fading, for any* $\alpha > 2$, $r_{\min} \in (0,1)$, $\mu \in (0,1)$, *and* $\delta \in (0, 1/2)$ *there exists*

$$b_3(n) \geq n^{-O\left(\log^{\delta-1/2}(n)\right)}$$

such that for any n, *node placement* $V(n)$ *with minimum separation* $r_{\min}$, *and permutation traffic matrix* $\lambda^{\mathsf{UC}}(n)$, *we have*

$$\rho^*(n) \geq \rho^{\mathsf{CMH}}(n) \geq b_3(n) d^{*3-\alpha}(n) n^{-1/2},$$

where

$$d^*(n) \triangleq \min\{h : V(n) \text{ is } \mu \text{ regular at resolution } h\}.$$

The same conclusion holds for slow fading with probability $1 - o(1)$ *as* $n \to \infty$.

Theorem 3.4 shows that if $V(n)$ is regular at resolution $d^*(n)$ then a per-node rate of at least $\rho^{\mathsf{CMH}}(n) \geq d^{*3-\alpha}(n) n^{-1/2-\beta(n)}$ is achievable, where, as before, the "loss" term $\beta(n)$ converges to zero as $n \to \infty$ at a rate arbitrarily close to $O\left(\log^{-1/2}(n)\right)$. The performance of the cooperative multi-hop scheme can intuitively be understood as follows. The scheme achieves cooperation on a scale of $d^2(n)$. This leads to a multi-antenna gain of order $d^2(n)$. On the other hand, communication is over a distance of order $d(n)$, leading to a power loss of order $d^{-\alpha}(n)$. Moreover, each source-destination pair at a distance of order $n^{1/2}$ must transmit its data over order $n^{1/2} d^{-1}(n)$ many hops, leading to a multi-hop loss of $n^{-1/2} d(n)$. Combining these three factors results in a per-node rate of $d^{3-\alpha}(n) n^{-1/2}$.

The next theorem shows that Theorem 3.4 is tight under adversarial node placement under a constraint on the regularity. The proof of the theorem is presented in Section 3.9.

Theorem 3.5. *Under either fast or slow fading, for any* $\alpha > 3$, *there exists* $b_4(n) = O\left(\log^6(n)\right)$, *such that for any* n, *and* $d^*(n)$, *there exists a node placement* $V(n)$ *with*

minimum separation $1/2$ *and* $1/2$*-regular at resolution* $d^*(n)$ *such that for* $\lambda^{\mathsf{UC}}(n)$ *chosen uniformly at random from the set of all permutation traffic matrices, we have*

$$\rho^*(n) \leq b_4(n) {d^*}^{3-\alpha}(n) n^{-1/2},$$

with probability $1 - o(1)$ *as* $n \to \infty$.

As an example, assume that

$$d^*(n) = n^\eta$$

for some $\eta \geq 0$. Then Theorem 3.4 shows that for any node placement of regularity $d^*(n)$ and $\alpha > 3$,

$$\rho^{\mathsf{CMH}}(n) \geq n^{(3-\alpha)\eta - 1/2 - \beta(n)},$$

where $\beta(n)$ converges to zero as $n \to \infty$ at a rate arbitrarily close to $O\big(\log^{-1/2}(n)\big)$. In other words

$$\lim_{n\to\infty} \frac{\log(\rho^{\mathsf{CMH}}(n))}{\log(n)} \geq (3-\alpha)\eta - 1/2.$$

Moreover, by Theorem 3.5 there exist node placements with same regularity such that for random permutation traffic with high probability $\rho^*(n)$ is (essentially) of the same order, in the sense that

$$\lim_{n\to\infty} \frac{\log(\rho^*(n))}{\log(n)} \leq (3-\alpha)\eta - 1/2.$$

In particular, for $\eta = 0$ (i.e., regular node placement), and for $\eta = \log\log(n)/\log(n)$ (i.e., random node placement), we obtain the order $n^{-1/2}$ scaling as expected. For $\eta = 1/2$ (i.e., completely irregular node placement), we obtain the order $n^{1-\alpha/2}$ scaling as in Theorems 3.1 and 3.3.

3.2 Hierarchical Relaying Scheme

This section describes the architecture of our hierarchical relaying scheme. On a high level, the construction of this scheme is as follows. Consider n nodes $V(n)$ placed

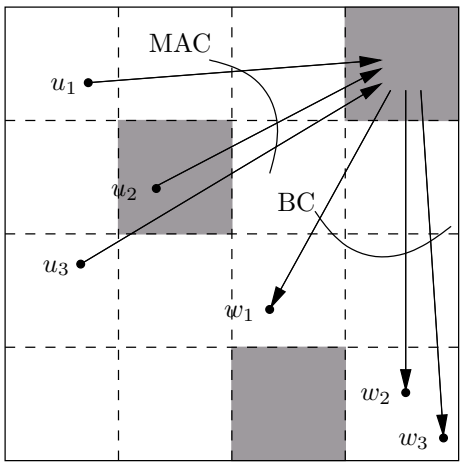

Figure 3-1: Sketch of one level of the hierarchical relaying scheme. Here $\{(u_i, w_i)\}_{i=1}^{3}$ are three source-destination pairs. Groups of source-destination pairs relay their traffic over dense subsquares, which contain a number of nodes proportional to their area (shaded). We time share between the different dense subsquares used as relays. Within each of these relay subsquares the scheme is used recursively to enable joint decoding and encoding at each relay.

arbitrarily on the square region $A(n)$ with a minimum separation $r_{\min}$. Divide $A(n)$ into subsquares of equal size. Call a subsquare *dense*, if it contains a number of nodes proportional to its area. For each source-destination pair, choose such a dense subsquare as a *relay*, over which it will transmit information (see Figure 3-1).

Consider now one such relay subsquare and the nodes that are transmitting information over it. If we assume for the moment that all the nodes within the same relay subsquare could cooperate then we would have a multiple access channel (MAC) between the source nodes and the relay subsquare, where each of the source nodes has one transmit antenna, and the relay subsquare (acting as one node) has many receive antennas. Between the relay subsquare and the destination nodes, we would have a broadcast channel (BC), where each destination node has one receive antenna, and the relay subsquare (acting again as one node) has many transmit antennas. The cooperation gain from using this kind of scheme arises from the use of multiple antennas for these multiple access and broadcast channels.

To actually enable this kind of cooperation at the relay subsquare, local communication within the relay subsquares is necessary. It can be shown that this local

communication problem is actually the same as the original problem, but at a smaller scale. Hence we can use the same scheme recursively to solve this subproblem. We terminate the recursion after several iterations, at which point we use simple time sharing to bootstrap the scheme.

The construction of the hierarchical relaying scheme is presented in detail in Section 3.2.1. A back-of-the-envelope calculation of the per-node rate it achieves is presented in Section 3.2.2. A detailed analysis of the hierarchical relaying scheme is presented in Sections 3.4 and 3.5.

3.2.1 Construction

Recall that

$$A(b) \triangleq [0, \sqrt{b}]^2$$

is the square region of area b. The scheme described here assumes that n nodes are placed arbitrarily in $A(n)$ with minimum separation $r_{\min} \in (0, 1)$. We want to find some rate, say ρ_0, that can be supported for all n source-destination pairs of a given permutation traffic matrix $\lambda^{\mathsf{UC}}(n)$. The scheme that is described below is "recursive" (and hence hierarchical) in the following sense. In order to achieve rate ρ_0 for n nodes in $A(n)$, it will use as a building block a scheme for supporting rate ρ_1 for a network of

$$n_1 \triangleq \frac{n}{2\gamma(n)}$$

nodes over $A(a_1)$ (square of area a_1) with

$$a_1 \triangleq \frac{n}{\gamma(n)}$$

for any permutation traffic matrix $\lambda^{\mathsf{UC}}(n_1)$ of n_1 nodes. Here the *branching factor* $\gamma(n)$ is a function such that $\gamma(n) \to \infty$ as $n \to \infty$. We will optimize over the choice of $\gamma(n)$ later. The same construction is used for the scheme over $A(a_1)$, and so on. In general, our scheme operates as follows at level $\ell \geq 0$ of the hierarchy (or recursion).

In order to achieve rate ρ_ℓ for any permutation traffic matrix $\lambda^{\mathsf{UC}}(n_\ell)$ over

$$n_\ell \triangleq \frac{n}{2^\ell \gamma^\ell(n)}$$

nodes in $A(a_\ell)$, with

$$a_\ell \triangleq \frac{n}{\gamma^\ell(n)},$$

use as a building block a scheme achieving rate $\rho_{\ell+1}$ over $n_{\ell+1}$ nodes in $A(a_{\ell+1})$ for any permutation traffic matrix $\lambda^{\mathsf{UC}}(n_{\ell+1})$. The recursion is terminated at some level $L(n)$ to be chosen later.

We now describe how the hierarchy is constructed between levels ℓ and $\ell+1$ for $0 \leq \ell < L(n)$. Each source-destination pair chooses some subsquare as a relay over which it transmits its message. This relaying of messages takes place in two phases — a *multiple access phase* and a *broadcast phase*. We first describe the selection of relay subsquares, then the operation of the network during the multiple access and broadcast phases, and finally the termination of the hierarchical construction.

Setting up Relays

Given n_ℓ nodes in $A(a_\ell)$, divide the square region $A(a_\ell)$ into $\gamma(n)$ equal sized subsquares. Denote them by $\{A_k(a_{\ell+1})\}_{k=1}^{\gamma(n)}$. Call a subsquare *dense* if it contains at least $n_\ell/2\gamma(n) = n_{\ell+1}$ nodes. In other words, a dense subsquare contains a number of nodes of at least a $1/2^{\ell+1}$ fraction of its area. We show that since the nodes in $A(a_\ell)$ have constant minimum separation $r_{\min}$, a subsquare can contain at most $O(a_{\ell+1})$ (i.e. $O(a_\ell/\gamma(n))$) nodes, and hence that there are at least $\Theta(2^{-\ell}\gamma(n))$ dense subsquares. Each source-destination pair chooses a dense subsquare such that both the source and the destination are at a distance $\Omega(\sqrt{a_{\ell+1}})$ from it. We call this dense subsquare the *relay* of this source-destination pair. We show that the relays can be chosen such that each relay subsquare has at most $n_{\ell+1}$ communication pairs that use it as relay, and we assume this worst case in the following discussion.

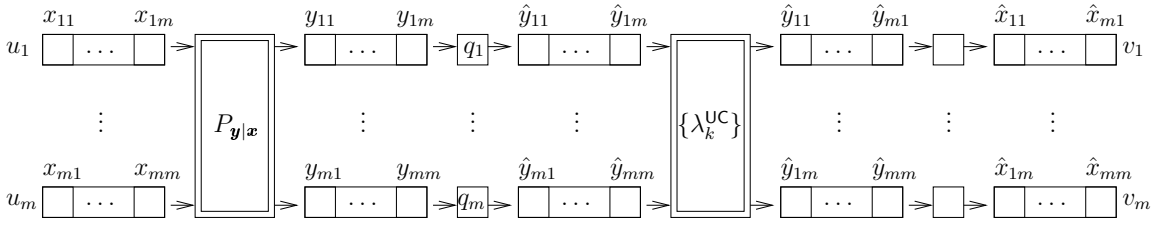

Figure 3-2: Description of the multiple access phase at level ℓ in the hierarchy with $m \triangleq n_{\ell+1}$. The first system block represents the wireless channel, connecting source nodes $\{u_i\}_{i=1}^{n_{\ell+1}}$ with relay nodes $\{v_i\}_{i=1}^{n_{\ell+1}}$. The second system block are quantizers $\{q_i\}_{i=1}^{n_{\ell+1}}$ used at the relay nodes. The third system block represents using $n_{\ell+1}$ times the communication scheme at level $\ell+1$ (organized as $n_{\ell+1}$ permutation traffic matrices $\{\lambda_k^{\mathsf{UC}}(n_{\ell+1})\}_{k=1}^{n_{\ell+1}}$) to "transpose" the matrix of quantized observations $\{\hat{y}_{ij}\}_{i,j=1}^{n_{\ell+1}}$. In other words, before the third system block, node v_1 has access to $\{\hat{y}_{1j}\}_{j=1}^{n_{\ell+1}}$, and after the third system block, node v_1 has access to $\{\hat{y}_{i1}\}_{i=1}^{n_{\ell+1}}$. The fourth system block are matched filters used at the relay nodes.

Multiple Access Phase

Source nodes that are assigned to the same (dense) relay subsquare send their messages simultaneously to that relay. We time share between the $\Theta(2^{-\ell}\gamma(n))$ different relay subsquares. If the nodes in the relay subsquare could cooperate, we would be dealing with a MAC with at most $n_{\ell+1}$ transmitters, each with one antenna, and one receiver with at least $n_{\ell+1}$ antennas. In order to achieve this cooperation, communication within the relay subsquare is necessary. To this end, each node in the relay subsquare quantizes its observations and then exchanges these quantized observations with the other nodes in the relay subsquare. This exchange is performed using a hierarchical (i.e., recursive) construction. For this recursive construction, assume that we have access to a communication scheme to transmit data according to a permutation traffic matrix $\lambda^{\mathsf{UC}}(n_{\ell+1})$ between $n_{\ell+1}$ nodes located in a square of area $a_{\ell+1}$. We now show how this scheme at scale $a_{\ell+1}$ can be used to construct a scheme for scale a_ℓ (see Figure 3-2).

Suppose there are $n_{\ell+1}$ source nodes $u_1, \ldots, u_{n_{\ell+1}}$ (located anywhere in $A(a_\ell)$) that relay their message over the $n_{\ell+1}$ relay nodes $v_1, \ldots, v_{n_{\ell+1}}$ (located in the same dense subsquare of area $a_{\ell+1}$). Each source node u_i divides its message bits into $n_{\ell+1}$ parts of equal length. Denote by x_{ij} the encoded part j of the message bits of node

u_i (x_{ij} is really a large sequence of channel symbols; to simplify the exposition, we shall, however, assume it is only a single symbol). The message parts corresponding to $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$ will be relayed over node v_j, as will become clear in the following. Sources $\{u_i\}_{i=1}^{n_{\ell+1}}$ transmit $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$ at time j for $j \in \{1, \dots n_{\ell+1}\}$.

Let y_{kj} be the observed channel output at relay v_k at time j. Note that y_{kj} depends only on channel inputs $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$. In order to decode the message parts corresponding to $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$ at relay node v_j, it needs to obtain the observations $\{y_{ij}\}_{i=1}^{n_{\ell+1}}$ from all other relay nodes. In other words, all relays need to exchange information. For this, each relay v_k quantizes its observation $\{y_{kj}\}_{j=1}^{n_{\ell+1}}$ at an appropriate rate K independent of n to obtain $\{\hat{y}_{kj}\}_{j=1}^{n_{\ell+1}}$. Quantized observation $\hat{y}_{kj}$ is to be sent from relay v_k to relay v_j. Thus, each of the $n_{\ell+1}$ relay nodes now has a message of size K for every other relay node.

This communication demand within the relay subsquare can be organized as $n_{\ell+1}$ permutation traffic matrices $\{\lambda_j^{\mathsf{UC}}(n_{\ell+1})\}_{j=1}^{n_{\ell+1}}$ between the $n_{\ell+1}$ relay nodes. Note that these relay nodes are located in the same square of area $a_{\ell+1}$. In other words, we are now faced with the original problem, but at smaller scale $a_{\ell+1}$. Therefore, using $n_{\ell+1}$ times the assumed scheme for transmitting according to a permutation traffic matrix for $n_{\ell+1}$ nodes in $A(a_{\ell+1})$, relay v_j can obtain all quantized observations $\{\hat{y}_{ij}\}_{i=1}^{n_{\ell+1}}$. Now v_j uses $n_{\ell+1}$ matched filters on $\{\hat{y}_{ij}\}_{i=1}^{n_{\ell+1}}$ to obtain estimates $\{\hat{x}_{ij}\}_{i=1}^{n_{\ell+1}}$ of $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$. In other words, each node v_j computes[1]

$$\hat{x}_{ij} = \sum_{k=1}^{n_{\ell+1}} \frac{h_{u_i,v_k}^{\dagger}[j]}{\sqrt{\sum_k |h_{u_i,v_k}[j]|^2}} \hat{y}_{kj}$$

for every $i \in \{1, \dots, n_{\ell+1}\}$. Using these estimates it then decodes the messages corresponding to $\{x_{ij}\}_{i=1}^{n_{\ell+1}}$.

[1]Note that, since we assume full CSI, node v_j has access to the channel gains $\{h_{u_i,v_k}[j]\}_{i,k}$ at any time $t \geq j$. In particular, this is the case at the time the matched filtering is performed.

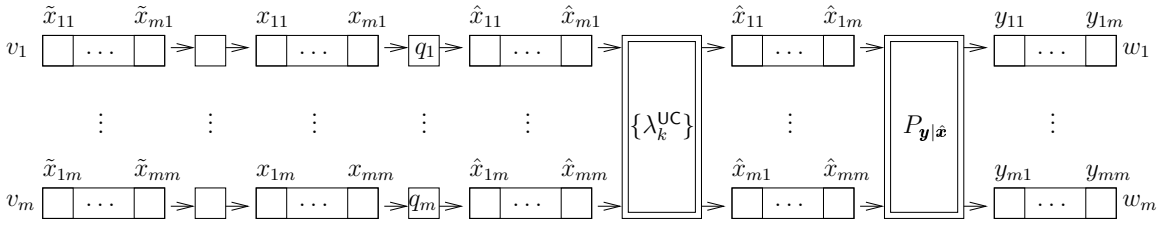

Figure 3-3: Description of the broadcast phase at level ℓ in the hierarchy with $m \triangleq n_{\ell+1}$. The first system block represents transmit beamforming at each of the relay nodes $\{v_i\}_{i=1}^{n_{\ell+1}}$. The second system block are quantizers $\{q_i\}_{i=1}^{n_{\ell+1}}$ used at the relay nodes. The third system block represents using $n_{\ell+1}$ times the communication scheme at level $\ell+1$ (organized as $n_{\ell+1}$ permutation traffic matrices $\{\lambda_k^{\mathsf{UC}}(n_{\ell+1})\}_{k=1}^{n_{\ell+1}}$) to "transpose" the matrix of quantized beamformed channel symbols $\{\hat{x}_{ij}\}_{i,j=1}^{n_{\ell+1}}$. In other words, before the third system block, node v_1 has access to $\{\hat{x}_{i1}\}_{i=1}^{n_{\ell+1}}$, and after the third system block, node v_1 has access to $\{\hat{x}_{1j}\}_{j=1}^{n_{\ell+1}}$. The fourth system block is the wireless channel, connecting relay nodes $\{v_i\}_{i=1}^{n_{\ell+1}}$ with destination nodes $\{w_i\}_{i=1}^{n_{\ell+1}}$.

Broadcast Phase

Nodes in the same relay subsquare then send their decoded messages simultaneously to the destination nodes corresponding to this relay. We time share between the different relay subsquares. If the nodes in the relay subsquare could cooperate, we would be dealing with a BC with one transmitter with at least $n_{\ell+1}$ antennas and with at most $n_{\ell+1}$ receivers, each with one antenna. In order to achieve this cooperation, a similar hierarchical construction as for the MAC phase is used. As in the MAC phase, assume that we have access to a scheme to transmit data according to a permutation traffic matrix $\lambda^{\mathsf{UC}}(n_{\ell+1})$ between $n_{\ell+1}$ nodes located in a square of area $a_{\ell+1}$. We again use this scheme at scale $a_{\ell+1}$ in the construction of the scheme for scale a_ℓ (see Figure 3-3).

Suppose there are $n_{\ell+1}$ relay nodes $v_1, \ldots, v_{n_{\ell+1}}$ (located in the same dense subsquare of area $a_{\ell+1}$) that relay traffic for $n_{\ell+1}$ destination nodes $w_1, \ldots, w_{n_{\ell+1}}$ (located anywhere in $A(a_\ell)$). Recall that at the end of the MAC phase, each relay node v_j has (assuming decoding was successful) access to parts j of the message bits of all source nodes $\{u_i\}_{i=1}^{n_{\ell+1}}$. Node v_j re-encodes these parts independently; call $\{\tilde{x}_{ij}\}_{i=1}^{n_{\ell+1}}$ the encoded channel symbols (as before, we assume $\tilde{x}_{ij}$ is only a single symbol to simplify exposition). Relay node v_j then performs transmit beamforming on $\{\tilde{x}_{ij}\}_{i=1}^{n_{\ell+1}}$ for the

$n_{\ell+1}$ transmit antennas of $\{v_k\}_{k=1}^{n_{\ell+1}}$ to be sent at time $T+j$ (for some appropriately chosen $T>0$ not depending on j). Call x_{kj} the resulting channel symbol to be sent from relay node v_k. Then[2]

$$x_{kj} = \sum_i \frac{h^\dagger_{v_k,w_i}[T+j]}{\sqrt{\sum_k |h_{v_k,w_i}[T+j]|^2}} \tilde{x}_{ij}.$$

In order to actually send this channel symbol, relay node v_k needs to obtain x_{kj} from node v_j. Thus, again all relay nodes need to exchange information.

To enable local cooperation within the relay subsquare, each relay node v_j quantizes its beamformed channel symbols $\{x_{kj}\}_{k=1}^{n_{\ell+1}}$ at an appropriate rate $K\log(n)$ with K independent of n to obtain $\{\hat{x}_{kj}\}_{k=1}^{n_{\ell+1}}$. Now, quantized value $\hat{x}_{kj}$ is sent from relay v_j to relay v_k. Thus, each of the $n_{\ell+1}$ relay nodes now has a message of size $K\log(n)$ for every other relay node.

This communication demand within the relay subsquare can be organized as $n_{\ell+1}$ permutation traffic matrices $\{\lambda_k^{\mathsf{UC}}(n_{\ell+1})\}_{k=1}^{n_{\ell+1}}$ between the $n_{\ell+1}$ relay nodes. Note that these relay nodes are located in the same square of area $a_{\ell+1}$. Hence, we are again faced with the original problem, but at smaller scale $a_{\ell+1}$. Using $n_{\ell+1}$ times the assumed scheme for transmitting according to a permutation traffic matrix for $n_{\ell+1}$ nodes in $A(a_{\ell+1})$, relay v_k can obtain all quantized beamformed channel symbols $\{\hat{x}_{kj}\}_{j=1}^{n_{\ell+1}}$. Now each v_k sends $\hat{x}_{kj}$ over the wireless channel at time instance $T+j$ (with T chosen to account for the preceding MAC phase and the local cooperation in the BC phase). Call y_{ij} the received channel output at destination node w_i at time instance $T+j$. Using y_{ij}, destination node w_i can now decode part j of the message bits of its source node u_i.

Spatial Re-Use and Termination of Recursion

The scheme performs appropriately weighted time division multiplexing among different levels $0 \le \ell \le L(n)$. Within any level $\ell \ge 1$, multiple regions of the original

[2]Note that, since we only assume causal CSI, relay node v_j does not actually have access to $\{h_{v_k,w_i}[T+j]\}_{k,i}$ at the time the beamforming is performed. This problem can, however, be circumvented. The details are provided in the proofs (see Lemma 3.10).

square $A(n)$ of area n are being operated in parallel. The details related to the effects of interference between different regions operating at the same level of hierarchy are discussed in the proofs.

The recursive construction terminates at some large enough level $L = L(n)$ (to be chosen later). At this scale, we have n_L nodes in area $A(a_L)$. A permutation traffic matrix at this level comprises n_L source-destination pairs. These transmissions are performed using simple time sharing. Again, multiple regions in the original square of area n at level L are active simultaneously.

3.2.2 Achievable Rates

Here we present a back-of-the-envelope calculation of the per-node rate $\rho^{\mathsf{HR}}(n)$ achievable with the hierarchical relaying scheme described in the previous section. The complete proof is stated in Section 3.5. We assume throughout that long block codes and corresponding optimal decoders are used for transmission.

Instead of computing the rate achieved by hierarchical relaying, it will be convenient to instead analyze its inverse, i.e., the time utilized for transmission of a single message bit from each source to its destination under a permutation traffic matrix $\lambda^{\mathsf{UC}}(n)$. Using the hierarchical relaying scheme, each message travels through L levels of the hierarchy. Call $\tau_\ell(n)$ the amount of time spent for the transmission of one message bit between each of the n_ℓ source-destination pairs at level ℓ in the hierarchy. We compute $\tau_\ell(n)$ recursively.

At any level $\ell \geq 1$, there are multiple regions of area a_ℓ operating at the same time. Due to the spatial re-use, each of these regions gets to transmit a constant fraction of time. It can be shown that the addition of interference due to this spatial re-use leads only to a constant loss in achievable rate. Hence the time required to send one message bit is only a constant factor higher than the one needed if region $A(a_\ell)$ is considered separately. Consider now one such region $A(a_\ell)$. By the time-sharing construction, only one of its $\Theta(2^{-\ell}\gamma(n))$ dense relay subsquares of area $a_{\ell+1}$ is active at any given moment. Hence the time required to operate all relay subsquares is a $\Theta(2^{-\ell}\gamma(n))$ factor higher than for just one relay subsquare separately. Consider now

one such relay subsquare, and assume $n_{\ell+1}$ source nodes in $A(a_\ell)$ communicate each $n_{\ell+1}$ message bits to their respective destination nodes through a MAC phase and BC phase with the help of the $n_{\ell+1}$ relay nodes in this relay subsquare of area $a_{\ell+1}$.

In the MAC phase, each of the $n_{\ell+1}$ sources simultaneously sends one bit to each of the $n_{\ell+1}$ relay nodes. The total time for this transmission is composed of two terms.

i) Transmission of $n_{\ell+1}$ message bits from each of the $n_{\ell+1}$ source nodes to equally many relay nodes. Since we time share between $\Theta(2^{-\ell}\gamma(n))$ relay subsquares, we can transmit with an average power constraint of $\Theta(2^{-\ell}\gamma(n))$ during the time a relay subsquare is active, and still satisfies the overall average power constraint of 1. With this "bursty" transmission strategy, we require a total of

$$O\left(n_{\ell+1}\frac{a_\ell^{\alpha/2}}{2^{-\ell}\gamma(n)n_{\ell+1}}\right) = O\big(n_{\ell+1}4^\ell\gamma^{\ell(1-\alpha/2)}(n)n^{\alpha/2-1}\big) \tag{3.1}$$

channel uses to transmit $n_{\ell+1}$ bits per source node. The terms on the left-hand side of (3.1) can be understood as follows: $n_{\ell+1}$ is the number of bits to be transmitted; $a_\ell^{\alpha/2}$ is the power loss since most nodes communicate over a distance of $\Theta(a_\ell^{1/2})$; $2^{-\ell}\gamma(n)$ is the average transmit power; $n_{\ell+1}$ is the multiple-antenna gain, since we have that many transmit and receive antennas.

ii) We show that constant rate quantization of the received observations at the relays is sufficient. Hence the $n_{\ell+1}$ bits for all sources generate $O(n_{\ell+1})$ transmissions at level $\ell+1$ of the hierarchy. Therefore,

$$O(n_{\ell+1}\tau_{\ell+1}(n)) \tag{3.2}$$

channel uses are needed to communicate all quantized observations to their respective relay nodes.

Combining (3.1) and (3.2), accounting for the factor $2^{-\ell}\gamma(n)$ loss due to time division between relay subsquares, we obtain that the transmission time for one message bit

from each source to the relay subsquare in the MAC phase at level ℓ is

$$\tau_\ell^{\mathsf{MAC}}(n) = O\Big(2^\ell \gamma^{1+\ell(1-\alpha/2)}(n) n^{\alpha/2-1} + \tau_{\ell+1}(n)\Big). \tag{3.3}$$

Next, we compute the number of channel uses per message bit received by the destination nodes in the BC phase. Similar to the MAC phase, each of the $n_{\ell+1}$ relay nodes has $n_{\ell+1}$ message bits out of which one bit is to be transmitted to each of the $n_{\ell+1}$ destination nodes. Since there are $n_{\ell+1}$ relay nodes, each destination node receives $n_{\ell+1}$ message bits. As before the required transmission time has two components.

i) Transmission of the encoded and quantized message bits from each of the $n_{\ell+1}$ relay nodes to all other relay nodes at level $\ell+1$ of the hierarchy. We show that each message bit results in $O\big((\ell+1)\log n\big)$ quantized bits. Therefore, $O\big(n_{\ell+1}(\ell+1)\log n\big)$ bits need to be transmitted from each relay node. This requires

$$O\big(n_{\ell+1}(\ell+1)\log(n)\tau_{\ell+1}(n)\big) \tag{3.4}$$

channel uses.

ii) Transmission of $n_{\ell+1}$ message bits from the relay nodes to each destination node. As before, we use bursty transmission with an average power constraint of $\Theta(2^{-\ell}\gamma(n))$ during the fraction $\Theta(2^\ell \gamma^{-1}(n))$ of time each relay subsquare is active (this satisfies the overall average power constraint of 1). Using this bursty strategy requires

$$O\left(n_{\ell+1}\frac{a_\ell^{\alpha/2}}{2^{-\ell}\gamma(n)n_{\ell+1}}\right) = O\big(n_{\ell+1}4^\ell\gamma^{\ell(1-\alpha/2)}(n)n^{\alpha/2-1}\big) \tag{3.5}$$

channel uses for transmission of $n_{\ell+1}$ bits per destination node. As in the MAC phase, $n_{\ell+1}$ in the left hand side of (3.5) can be understood as the number of bits to be transmitted, $a_\ell^{\alpha/2}$ as the power loss for communicating over distance $\Theta(a_\ell^{1/2})$, $2^{-\ell}\gamma(n)$ as the average transmit power, and $n_{\ell+1}$ as the multiple-

antenna gain.

Combining (3.4) and (3.5), accounting for a factor $2^{-\ell}\gamma(n)$ loss due to time division between relay subsquares, the transmission time for one message bit from the relays to each destination node in the BC phase at level ℓ is

$$\tau_\ell^{\mathsf{BC}}(n) = O\Big(2^\ell \gamma^{1+\ell(1-\alpha/2)}(n) n^{\alpha/2-1} + (\ell+1)\log(n)\tau_{\ell+1}(n)\Big). \tag{3.6}$$

From (3.3) and (3.6), we obtain the following recursion

$$\begin{aligned}
\tau_\ell(n) &= \tau_\ell^{\mathsf{MAC}}(n) + \tau_\ell^{\mathsf{BC}}(n) \\
&= O\Big(2^\ell \gamma^{1+\ell(1-\alpha/2)}(n) n^{\alpha/2-1} + (\ell+1)\log(n)\tau_{\ell+1}(n)\Big) \\
&= O\Big(2^L \gamma(n) n^{\alpha/2-1} + L\log(n)\tau_{\ell+1}(n)\Big),
\end{aligned} \tag{3.7}$$

where we have used $\alpha > 2$. This recursion holds for all $0 \le \ell < L$. At level L, we time share among n_L nodes in region $A(a_L)$ with a permutation traffic matrix $\lambda^{\mathsf{UC}}(n_L)$. Each of the n_L source-destination pairs uses the wireless channel for $1/n_L$ fraction of the time at power $O(n_L)$, satisfying the average power constraint. This achieves a rate of at least $\Omega(a_L^{-\alpha/2})$ between any source-destination pair. Equivalently

$$\begin{aligned}
\tau_L(n) &= O(a_L^{\alpha/2}) \\
&= O\big(n^{\alpha/2}\gamma^{-L\alpha/2}(n)\big) \\
&= O\big(n^{\alpha/2}\gamma^{-L}(n)\big).
\end{aligned} \tag{3.8}$$

Using the recursion (3.7) L times and combining with (3.8), we have

$$\begin{aligned}
\tau_0(n) &= O\big(n^{\alpha/2-1}2^L\gamma(n) + L\log(n)\tau_1(n)\big) \\
&= \dots \\
&= O\Big(n^{\alpha/2-1}\big(L\log(n)\big)^L 2^L\gamma(n) + \big(L\log(n)\big)^L\tau_L(n)\Big) \\
&= O\Big(n^{\alpha/2-1}\big(L\log(n)\big)^L\big(2^L\gamma(n) + n\gamma^{-L}(n)\big)\Big).
\end{aligned} \tag{3.9}$$

The term

$$\big(L\log(n)\big)^L\big(2^L\gamma(n)+n\gamma^{-L}(n)\big) \tag{3.10}$$

is the "loss" factor over the desired order $n^{\alpha/2-1}$ scaling, and we now choose the branching factor $\gamma(n)$ and the hierarchy depth $L \triangleq L(n)$ to make it small. Fix a $\delta \in (0, 1/2)$ and set[3]

$$L(n) \triangleq \log^{1/2-\delta}(n),$$
$$\gamma(n) \triangleq n^{1/(L(n)+1)}.$$

With this

$$\big(L(n)\log(n)\big)^{L(n)} \leq n^{2\log^{-1/2-\delta}(n)\log\log(n)},$$
$$2^{L(n)}\gamma(n) \leq n^{\log^{-1/2-\delta}(n)+\log^{\delta-1/2}(n)},$$
$$n\gamma^{-L(n)}(n) \leq n^{\log^{\delta-1/2}(n)}.$$

Since $\delta > 0$, the $n^{\log^{\delta-1/2}(n)}$ term dominates in (3.9), and we obtain

$$\tau_0(n) \leq \tilde{b}(n)n^{\alpha/2-1},$$

where

$$\tilde{b}(n) \leq n^{O(\log^{\delta-1/2}(n))}.$$

Hence the per-node rate of the hierarchical relaying scheme is lower bounded as

$$\rho^{\mathsf{HR}}(n) = 1/\tau_0(n) \geq b(n)n^{1-\alpha/2},$$

with

$$b(n) \geq n^{-O(\log^{\delta-1/2}(n))}.$$

[3]There are several choices of $L(n)$ and $\gamma(n)$ that result in the "loss" factor (3.10) to be of order $n^{o(1)}$. The choice here is convenient for the slow fading case discussed in detail in Chapter 3.5.2. However, other choices are possible as well.

Note that to minimize the loss term, we should choose $\delta > 0$ to be small.

3.3 Cooperative Multi-Hop Scheme

In this section, we provide a brief description of the cooperative multi-hop scheme. The details of the construction and the analysis of its performance can be found in Section 3.8.

Recall that a node placement $V(n)$ is μ-regular at resolution $d(n)$ if every square $[id(n), (i+1)d(n)] \times [jd(n), (j+1)d(n)]$ for some $i, j \in \mathbb{N}$ contains at least $\mu d^2(n)$ nodes. Given such a node placement $V(n)$, divide it into squares of sidelength $d(n)$. Consider four adjacent squares, combined into a bigger square of sidelength $2d(n)$. By the regularity assumption on $V(n)$, this bigger square contains at least $4\mu d^2(n)$ nodes. Hence we can apply the hierarchical relaying scheme introduced in Section 3.2 to support any permutation traffic within this bigger square at a per-node rate of

$$b(n)(d^2(n))^{1-\alpha/2} = b(n)d^{2-\alpha}(n),$$

where $b(n)$ is essentially of order $n^{-\log^{-1/2}(n)}$. By properly choosing the permutation traffic matrices within every possible such bigger square of sidelength $2d(n)$ and with appropriate spatial re-use, nodes in each square of sidelength $d(n)$ can communicate with neighboring squares at a sum rate of

$$d^2(n)b(n)d^{2-\alpha}(n) = b(n)d^{4-\alpha}(n).$$

We now construct a graph with $n/d^2(n)$ vertices, each corresponding to one square of sidelength d^n in $A(n)$. Nodes corresponding to neighboring squares are connected by an edge with edge capacity

$$b(n)d^{4-\alpha}(n).$$

The resulting graph is depicted in Figure 3-4. Note that with the above communication procedure if messages can be routed over this graph then the same messages can

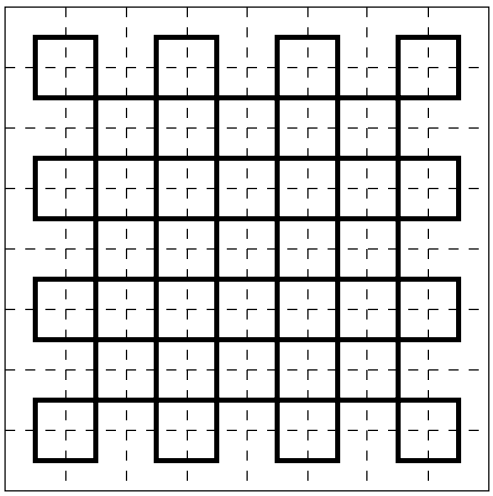

Figure 3-4: Graph (in bold) resulting from the construction of the cooperative multi-hop scheme. The entire square has sidelength $\sqrt{n}$, and the dashed squares have sidelength $d(n)$. Each (bold) edge in the graph corresponds to using the hierarchical relaying scheme between the nodes in the adjacent squares of sidelength $d(n)$.

be reliably communicated over the wireless network.

Now, to send a message from a source node in $V(n)$ to its destination node, we first locate the squares of sidelength $d(n)$ they are located in. We then route the message over the edges of the graph constructed above in a multi-hop fashion. By the construction of the graph, each such edge is implemented using the hierarchical relaying scheme. In other words, we perform multi-hop communication over distance $\sqrt{n}$ with hop length $d(n)$, and each such hop is implemented using hierarchical relaying over distance $d(n)$. Since each edge in the graph has a capacity of $b(n)d^{4-\alpha}(n)$ and has to support roughly $n^{1/2}d(n)$ source-destination pairs, we obtain a per-node rate of

$$\begin{aligned}\rho^{\mathsf{CMH}}(n) &\geq b(n)d^{4-\alpha}(n)n^{-1/2}d^{-1}(n)\\ &= b(n)d^{3-\alpha}(n)n^{-1/2}\end{aligned}$$

per source-destination pair.

3.4 Analysis of the Hierarchical Relaying Scheme

In this section, we analyze in detail the hierarchical relaying scheme. Throughout Sections 3.4.1 to 3.4.3, we consider communication at level ℓ, $0 \leq \ell < L = L(n)$, of the hierarchy. All constants K_i are independent of ℓ.

Recall that at level ℓ in the hierarchy, we have a square region $A(a_\ell)$ of area

$$a_\ell \triangleq \frac{n}{\gamma^\ell(n)}$$

containing (at least)

$$n_\ell \triangleq \frac{n}{2^\ell \gamma^\ell(n)}$$

nodes $V(n_\ell)$. We divide $A(a_\ell)$ into $\gamma(n)$ subsquares of area $a_{\ell+1}$. Recall that a subsquare of area $a_{\ell+1}$ in level ℓ of the hierarchy is called dense if it contains at least $n_{\ell+1}$ nodes. We impose a power constraint of $P_\ell(n) = \Theta(2^{-\ell}\gamma(n))$ during the time any particular relay subsquare is active. Since we time share between $\Theta(2^{-\ell}\gamma(n))$ relay subsquares, this satisfies the overall average power constraint (by choosing constants appropriately).

Since other regions of area a_ℓ are active at the same time as the one under consideration, we have to deal with interference. To this end, we consider a slightly more general noise model that includes the experienced interference at the relay subsquares. More precisely, we assume that, for all $u \in V(n_\ell)$, the additive noise term $\{z_u[t]\}_t$ is independent of the signal $\{x_u[t]\}_t$ and of the channel gains $\{h_{u,v}[t]\}_{v,t}$; that the noise term is stationary and ergodic across time t, but with arbitrary dependence across nodes u; and that the noise has zero mean and bounded power N_0 independent of n. Note that we do not require the additive noise term to be Gaussian. In the above, N_0 accounts for both noise (which has power 1 in the original model) as well as interference. We show in Section 3.5 that these assumptions are valid.

Recall the following choice of $\gamma(n)$ and $L(n)$:

$$\begin{aligned} L(n) &\triangleq \log^{1/2-\delta}(n), \\ \gamma(n) &\triangleq n^{1/(L(n)+1)}, \end{aligned} \tag{3.11}$$

with $\delta \in (0, 1/2)$ independent of n. This choice satisfies

$$\begin{aligned} \gamma(n) &\leq \gamma(\tilde{n}) \qquad \text{if } n \leq \tilde{n}, \\ \gamma^{L(n)}(n) &\leq n \qquad\quad \text{for all } n, \\ 2^{-L(n)}\gamma(n) &\to \infty \qquad \text{as } n \to \infty, \end{aligned} \tag{3.12}$$

The first condition in (3.12) implies that the number of subsquares $\gamma(n)$ we divide $A(n)$ into increases in n. The second condition implies that the subsquare area $a_{L(n)}$ at the last level of the hierarchy is bigger than 1. As we shall see, the third condition implies that the number of dense subsquares at the last level of the hierarchy (and hence at every level) grows unbounded as $n \to \infty$ (see Lemma 3.6 below).

Throughout Section 3.4, we consider the fast fading channel model. Slow fading is discussed in Section 3.5.2.

3.4.1 Setting up Relays

The first lemma states that the minimum-separation requirement $r_{\min} \in (0, 1)$ implies that a constant fraction of subsquares must be dense. We point out that this is the only consequence of the minimum-separation requirement used to prove Theorem 3.1. Thus Theorem 3.1 remains valid if we just assume that Lemma 3.6 below holds directly. See also Section 3.10.4 for further details.

Lemma 3.6. *For any $V(n_\ell) \subset A(a_\ell)$ with $|V(n_\ell)| \geq n_\ell$ and with minimum separation $r_{\min} \in (0, 1)$, each of its subsquares of area $a_{\ell+1}$ contains at most $K_1 a_\ell/\gamma(n)$ nodes, and there are at least $K_2 2^{-\ell}\gamma(n)$ dense subsquares, for some constants K_1 and K_2.*

Proof. Put a circle of radius $r_{\min}/2$ around each node. By the minimum-separation requirement, these circles do not intersect. Each node covers an area of $\pi r_{\min}^2/4$.

Increasing the sidelength of each subsquare by $r_{\min}$, this provides a total area of

$$\left(\sqrt{a_\ell/\gamma(n)} + r_{\min}\right)^2 \leq \frac{a_\ell}{\gamma(n)}(1 + r_{\min})^2$$

in which the circles around these nodes are packed. Here we have used that $\gamma^{\ell+1}(n) \leq n$ by (3.12), and therefore

$$\gamma(n) \leq n/\gamma^\ell(n) = a_\ell.$$

Hence there can be at most $K_1 a_\ell/\gamma(n)$ nodes per subsquare with

$$K_1 \triangleq 4\frac{(1 + r_{\min})^2}{\pi r_{\min}^2}.$$

Note that, since $r_{\min} < 1$, we have $K_1 > 1$.

Let $d(n_\ell)$ be the number of dense subsquares in $A(a_\ell)$, and therefore $\gamma(n) - d(n_\ell)$ is the number of subsquares that are not dense. By the argument in the last paragraph, each dense subsquare contains at most $K_1 a_\ell/\gamma(n)$ nodes, and those subsquares that are not dense contain less than $n_{\ell+1}$ nodes by the definition of dense subsquares. Hence $d(n_\ell)$ must satisfy

$$d(n_\ell)K_1 a_\ell/\gamma(n) + \big(\gamma(n) - d(n_\ell)\big)n_{\ell+1} \geq |V(n_\ell)| \geq n_\ell.$$

Thus, using $a_\ell = 2^\ell n_\ell$, $n_{\ell+1} = n_\ell/2\gamma(n)$, we have

$$d(n_\ell)K_1 2^\ell + (\gamma(n) - d(n_\ell))/2 \geq \gamma(n).$$

As $K_1 > 1$, this yields

$$d(n_\ell) \geq \frac{1 - 1/2}{K_1 2^\ell - 1/2}\gamma(n) \geq \frac{2^{-\ell}}{2K_1}\gamma(n) = K_2 2^{-\ell}\gamma(n),$$

with

$$K_2 \triangleq \frac{1}{2K_1}.$$

□

Consider $V(n_\ell) \subset A(a_\ell)$ with $|V(n_\ell)| \geq n_\ell$, and choose arbitrary $K_2 2^{-\ell}\gamma(n)$ dense subsquares of area $a_{\ell+1}$ (as guaranteed by Lemma 3.6). Call those subsquares $\{A_k(a_{\ell+1})\}_{k=1}^{K_2 2^{-\ell}\gamma(n)}$. For each source-destination pair, we now select one such dense subsquare to relay traffic over. To avoid bottlenecks, this selection has to be done such that all relay subsquares carry approximately the same amount of traffic. Moreover, for technical reasons, the distances from the source and the destination to the relay subsquare cannot be too small.

Formally, the selection of relay subsquares can be described by the *schedules* $S \in \{0,1\}^{n_\ell \times K_2 2^{-\ell}\gamma(n)}$ with $s_{u,k} = 1$ if source node u relays traffic over dense subsquare k, and $\widetilde{S} \in \{0,1\}^{K_2 2^{-\ell}\gamma(n) \times n_\ell}$ with $\tilde{s}_{k,w} = 1$ if destination node w receives traffic from dense subsquare k. With slight abuse of notation, let $r_{u,A_k(a_{\ell+1})}$ be the distance between node $u \in V(n_\ell)$ and the closest point in $A_k(a_{\ell+1})$, i.e.,

$$r_{u,A_k(a_{\ell+1})} \triangleq \min_{v \in A_k(a_{\ell+1})} r_{u,v}. \tag{3.13}$$

Define the sets

$$\begin{aligned}\mathcal{S}(n_\ell) \triangleq \Big\{ S \in \{0,1\}^{n_\ell \times K_2 2^{-\ell}\gamma(n)} :& 0 \leq \textstyle\sum_{u=1}^{n_\ell} s_{u,k} \leq n_{\ell+1} \ \forall k, \\ & 0 \leq \textstyle\sum_{k=1}^{K_2 2^{-\ell}\gamma(n)} s_{u,k} \leq 1 \ \forall u, \\ & s_{u,k} = 1 \text{ implies } r_{u,A_k(a_{\ell+1})} \geq \sqrt{2a_{\ell+1}} \ \forall u,k \Big\}\end{aligned} \tag{3.14}$$

and

$$\widetilde{\mathcal{S}}(n_\ell) \triangleq \big\{ \widetilde{S} \in \{0,1\}^{K_2 2^{-\ell}\gamma(n) \times n_\ell} : \widetilde{S}^T \in \mathcal{S}(n_\ell) \big\}.$$

The sets $\mathcal{S}(n_\ell)$ and $\widetilde{\mathcal{S}}(n_\ell)$ are the collection of schedules satisfying the conditions mentioned in the last paragraph. More precisely, the first condition in (3.14) ensures that at most $n_{\ell+1}$ source-destination pairs relay over the same dense subsquare, the second condition ensures that each source-destination pair chooses at most one relay subsquare, and the third condition ensures that sources and destinations are at least at distance $\sqrt{2a_{\ell+1}}$ from the chosen relay subsquare.

Next, we prove that any node placement that satisfies Lemma 3.6 allows for a de-

composition of any permutation traffic matrix $\lambda^{\mathsf{UC}}(n_\ell)$ into a small number of schedules belonging to $\mathcal{S}(n_\ell)$ and $\widetilde{\mathcal{S}}(n_\ell)$.

Lemma 3.7. *There exists K_3 such that for all n large enough (independent of ℓ), and every permutation traffic matrix $\lambda^{\mathsf{UC}}(n_\ell) \in \{0,1\}^{n_\ell \times n_\ell}$ we can find $K_3 2^\ell$ schedules $\{S^{(i)}(n_\ell)\}_{i=1}^{K_3 2^\ell} \subset \mathcal{S}(n_\ell)$, $\{\widetilde{S}^{(i)}(n_\ell)\}_{i=1}^{K_3 2^\ell} \subset \widetilde{\mathcal{S}}(n_\ell)$ satisfying*

$$\lambda^{\mathsf{UC}}(n_\ell) = \sum_{i=1}^{K_3 2^\ell} S^{(i)}(n_\ell)\widetilde{S}^{(i)}(n_\ell),$$

for some constant K_3.

Proof. Pick an arbitrary source-destination pair in $\lambda^{\mathsf{UC}}(n_\ell)$, and consider the subsquares containing the source and the destination node. Since each subsquare has side length $\sqrt{a_{\ell+1}}$, there are at most 50 subsquares at distance less than $\sqrt{2a_{\ell+1}}$ from either of those two subsquares. As $2^{-L(n)}\gamma(n) \to \infty$ as $n \to \infty$ by (3.12), there exists K (independent of ℓ) such that for $n \geq K$ we have $50 \leq K_2 2^{-\ell-1}\gamma(n)$. Since there are at least $K_2 2^{-\ell}\gamma(n)$ dense subsquares by Lemma 3.6, there must exist at least $K_2 2^{-\ell-1}\gamma(n)$ dense subsquares that are at distance at least $\sqrt{2a_{\ell+1}}$ from both the subsquares containing the source and the destination node.

In order to construct a decomposition of $\lambda^{\mathsf{UC}}(n_\ell)$, we use the following procedure. Sequentially, each of the n_ℓ source-destination pairs chooses one of the (at least) $K_2 2^{-\ell-1}\gamma(n)$ dense subsquares at distance at least $\sqrt{2a_{\ell+1}}$ that has not already been chosen by $n_{\ell+1}$ other pairs. If any source-destination pair cannot select such a subsquare, then stop the procedure and use the source-destination pairs matched with dense subsquares so far to define matrices $S^{(1)}(n_\ell)$ and $\widetilde{S}^{(1)}(n_\ell)$. Now, remove all the matched source-destination pairs, forget that dense subsquares were matched to any source-destination pair and repeat the above procedure, going through the remaining source-destination pairs.

Let

$$K_3 \triangleq 4/K_2.$$

We claim that by repeating this process of generating matrices $S^{(i)}(n_\ell)$ and $\widetilde{S}^{(i)}(n_\ell)$,

we can match all source-destination pairs to some dense subsquare with at most $K_3 2^\ell$ such matrices. Indeed, a new pair of matrices is generated only when a source-destination pair cannot be matched to any of its available (at least) $K_2 2^{-\ell-1}\gamma(n)$ dense subsquares. If this happens, all these dense subsquares are matched by $n_{\ell+1} = n_\ell/2\gamma(n)$ pairs. Hence at least $K_2 2^{-\ell-2} n_\ell$ source-destination pairs are matched in each "round". Since there are n_ℓ total pairs, we need at most

$$\frac{n_\ell}{K_2 2^{-\ell-2} n_\ell} = K_3 2^\ell$$

matrices $S^{(i)}(n_\ell)$ and $\widetilde{S}^{(i)}(n_\ell)$. □

For a permutation traffic matrix $\lambda^{\mathsf{UC}}(n_\ell)$, communication proceeds as follows. Write

$$\lambda^{\mathsf{UC}}(n_\ell) = \sum_{i=1}^{K_3 2^\ell} S^{(i)}(n_\ell)\widetilde{S}^{(i)}(n_\ell)$$

as in Lemma 3.7. Split time into $K_3 2^\ell$ equal length time slots. In slot i, we use $S^{(i)}(n_\ell)\widetilde{S}^{(i)}(n_\ell)$ as our traffic matrix. Consider without loss of generality $i = 1$ in the following. Write

$$S^{(1)}(n_\ell)\widetilde{S}^{(1)}(n_\ell) = \sum_{k=1}^{K_2 2^{-\ell}\gamma(n)} S^{(1,k)}(n_{\ell+1})\widetilde{S}^{(1,k)}(n_{\ell+1}),$$

where $S^{(1,k)}(n_{\ell+1})\widetilde{S}^{(1,k)}(n_{\ell+1})$ is the traffic relayed over the dense subsquare $A_k(a_{\ell+1})$. We time share between the schedules for $k \in \{1, \ldots, K_2 2^{-\ell}\gamma(n)\}$. Consider now any such k. In the worst case, there are exactly $n_{\ell+1}$ communication pairs to be relayed over $A_k(a_{\ell+1})$, and the relay subsquare $A_k(a_{\ell+1})$ contains exactly $n_{\ell+1}$ nodes. We shall assume this worst case in the following.

We focus on transmission according to the traffic matrix $S^{(1,1)}(n_{\ell+1})\widetilde{S}^{(1,1)}(n_{\ell+1})$. Let $V(n_{\ell+1})$ be the nodes in $A_1(a_{\ell+1})$, and let $U(n_{\ell+1})$ and $W(n_{\ell+1})$ be the source and destination nodes of $S^{(1,1)}(n_{\ell+1})\widetilde{S}^{(1,1)}(n_{\ell+1})$, respectively. In other words, the source nodes $U(n_{\ell+1})$ communicate to their respective destination nodes $W(n_{\ell+1})$ using the nodes $V(n_{\ell+1})$ as relays.

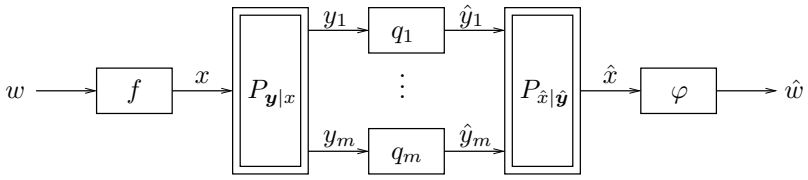

Figure 3-5: Sketch of the quantized channel. f and φ are the channel encoder and decoder, respectively; $\{q_k\}_{k=1}^m$ are quantizers; $P_{\boldsymbol{y}|x}$ and $P_{\hat{x}|\hat{\boldsymbol{y}}}$ represent stationary ergodic channels with the indicated marginal distributions.

3.4.2 Multiple Access Phase

Each source node in $U(n_{\ell+1})$ splits its message into $n_{\ell+1}$ equal length parts. Part j at every node $u \in U(n_{\ell+1})$ is to be relayed over the j-th node in $V(n_{\ell+1})$. Each part is separately encoded at the source and separately decoded at the destination. After the source nodes are done transmitting their messages, the nodes in the relay subsquare quantize their (sampled) observations corresponding to part j and communicate the quantized values to the j-th node in the relay subsquare. This node then decodes the j-th message parts of all source nodes. Note that this induces a uniform traffic pattern between the nodes in the relay subsquare, i.e., every node needs to transmit quantized observations to every other node. While this traffic pattern does not correspond to a permutation traffic matrix, it can be written as a sum of $n_{\ell+1}$ permutation traffic matrices. A $1/n_{\ell+1}$ fraction of the traffic within the relay subsquare is transmitted according to each of these permutation traffic matrices. This setup is depicted in Figure 3-2 in Section 3.2.1.

Assuming for the moment that we have a scheme to send the quantized observations to the dedicated node in the relay subsquare, the traffic matrix $S^{(1,1)}(n_{\ell+1})$ between $U(n_{\ell+1})$ and $V(n_{\ell+1})$ describes then a MAC with $n_{\ell+1}$ transmitters, each with one antenna, and one receiver with $n_{\ell+1}$ antennas. We call this the *MAC induced by* $S^{(1,1)}(n_{\ell+1})$ in the following. Before we analyze the rate achievable over this induced MAC, we need an auxiliary result on quantized channels.

Consider the quantized channel in Figure 3-5. Here, f is the channel encoder, φ the channel decoder, $\{q_k\}_{k=1}^m$ quantizers. All these have to be chosen. $P_{\boldsymbol{y}|x}$ and $P_{\hat{x}|\hat{\boldsymbol{y}}}$,

on the other hand, represent fixed stationary ergodic channels with the indicated marginal distributions. We call R the rate of the channel code (f, φ) and $\{R_k\}_{k=1}^m$ the rates of quantizers $\{q_k\}_{k=1}^m$.

Lemma 3.8. *If there exist distributions P_x and $\{P_{\hat{y}_k|y_k}\}_{k=1}^m$ such that $R < I(x;\hat{x})$ and $R_k > I(y_k;\hat{y}_k)$, $\forall k$, then $\left(R, \{R_k\}_{k=1}^m\right)$ is achievable over the quantized channel.*

Proof. The proof follows from a simple extension of Theorem 1 in Appendix II of [38]. □

Lemma 3.9. *Let the additive noise $\{z_v\}_{v\in V(n_{\ell+1})}$ be uncorrelated (over v). For the MAC induced by $S^{(1,1)}(n_{\ell+1})$ with per-node average power constraint $P_\ell(n) \le n_{\ell+1}^{-1} a_\ell^{\alpha/2}$, a rate of*

$$\rho_\ell^{\mathsf{MAC}}(n) \ge K_4 P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}$$

per source node is achievable, and the number of bits required at each relay node to quantize the observations is at most K_5 bits per $n_{\ell+1}$ total message bits[4] sent by the source nodes, for some constants K_4 and K_5.

Proof. The source nodes send signals with a power of (essentially) $n_{\ell+1}^{-1} a_\ell^{\alpha/2}$ for a fraction $P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2} \le 1$ of time and are silent for the remaining time. To ensure that interference is uniform, the time slots during which the nodes send signals are chosen randomly as follows. Generate independently for each region $A(a_\ell)$ a Bernoulli process $\{B[t]\}_{t\in\mathbb{N}}$ with parameter $P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}/(1+\eta) \le 1$ for some small $\eta > 0$. The nodes in $A(a_\ell)$ are active whenever $B[t] = 1$ and remain silent otherwise. Since the blocklength of the codes used is assumed to be large, this satisfies the average power constraint of $P_\ell(n)$ with high probability for any $\eta > 0$. Since we are interested only in the scaling of capacity, we ignore the additional $1/(1+\eta)$ term in the following to simplify notation. Clearly, we only need to consider the fraction of time during which $B[t] = 1$.

Let $\boldsymbol{y}$ be the received vector at the relay subsquare, $\hat{\boldsymbol{y}}$ the (componentwise) quan-

[4]Total message bits refers to the sum of all message bits transmitted by the $n_{\ell+1}$ source nodes.

tized observations. We use a matched filter at the relay subsquare, i.e.,

$$\hat{x}_u = \frac{\boldsymbol{h}_u^\dagger}{\|\boldsymbol{h}_u\|}\hat{\boldsymbol{y}},$$

where column vector $\boldsymbol{h}_u = \{h_{u,v}\}_{v\in V(n_{\ell+1})}$ are the channel gains between node $u \in U(n_{\ell+1})$ and the nodes in the relay subsquare $V(n_{\ell+1})$. The use of a matched filter is possible since we assume full CSI is available at all the nodes.

We now use Lemma 3.8 to show that we can design quantizers $\{q_v\}_{v\in V(n_{\ell+1})}$ of constant rate and achieve a per-node communication rate of at least $K_4 P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}$. The first channel in Lemma 3.8 (see Figure 3-5) will correspond to the wireless channel between a source node u and its relay subsquare $V(n_{\ell+1})$. The second "channel" in Lemma 3.8 will correspond to the matched filter used at the relay subsquare. To apply Lemma 3.8, we need to find a distribution for x_u and for $\hat{y}_v|y_v$. Define

$$\tilde{r}_u \triangleq r_{u,A_1(a_{\ell+1})}/\sqrt{2a_\ell} \leq 1$$

with $r_{u,A_1(a_{\ell+1})}$ as in (3.13), to be the normalized distance of the source node $u \in U(n_{\ell+1})$ to the relay subsquare $A_1(a_{\ell+1})$. For each $u \in U(n_{\ell+1})$, let

$$x_u \sim \mathcal{N}_{\mathbb{C}}(0, \tilde{r}_u^\alpha n_{\ell+1}^{-1} a_\ell^{\alpha/2})$$

independent of $x_{\tilde{u}}$ for $u \neq \tilde{u}$, and let $\hat{y}_v = y_v + \tilde{z}_v$ for $\tilde{z}_v \sim \mathcal{N}_{\mathbb{C}}(0, \Delta^2)$ independent of $\boldsymbol{y}$ and for some $\Delta^2 > 0$. Note that the channel input x_u has power that depends on the normalized distance $\tilde{r}_u$ (i.e., only nodes $u \in U(n_{\ell+1})$ that are at maximal distance $\sqrt{2a_\ell}$ from the relay subsquare transmit at full available power). This is to ensure that all signals are received at roughly the same strength by the relays.

We proceed by computing $I(y_v; \hat{y}_v|\{h_{\tilde{u},\tilde{v}}\})$ and $I(x_u; \hat{x}_u|\{h_{\tilde{u},\tilde{v}}\})$ as required in Lemma 3.8 (the conditioning on $\{h_{\tilde{u},\tilde{v}}\}$ being due to the availability of full CSI). Note first that by construction of $S^{(1,1)}(n_{\ell+1})$ (see (3.14)), we have for $u \in U(n_{\ell+1})$ and $v \in V(n_{\ell+1})$

$$r_{u,A_1(a_{\ell+1})} \leq r_{u,v} \leq 2r_{u,A_1(a_{\ell+1})},$$

and hence

$$\frac{1}{2\sqrt{2a_\ell}} \leq \frac{\tilde{r}_u}{r_{u,v}} \leq \frac{1}{\sqrt{2a_\ell}}. \tag{3.15}$$

From this, and since $|h_{u,v}|^2 = r_{u,v}^{-\alpha}$, we obtain

$$\begin{aligned} 2^{-3\alpha/2} a_\ell^{-\alpha/2} \leq |h_{u,v}|^2 \tilde{r}_u^\alpha \leq 2^{-\alpha/2} a_\ell^{-\alpha/2}, \\ 2^{-3\alpha/2} n_{\ell+1} a_\ell^{-\alpha/2} \leq \|\boldsymbol{h}_u\|^2 \tilde{r}_u^\alpha \leq 2^{-\alpha/2} n_{\ell+1} a_\ell^{-\alpha/2}. \end{aligned} \tag{3.16}$$

We start by computing $I(y_v; \hat{y}_v | \{h_{\tilde{u},\tilde{v}}\})$. We have

$$\hat{y}_v = \sum_{u \in U(n_{\ell+1})} h_{u,v} x_u + z_v + z_{\tilde{v}},$$

and hence $\hat{y}_v$ has mean zero and variance

$$\begin{aligned} \mathbb{E}(|\hat{y}_v|^2) &= \sum_{u \in U(n_{\ell+1})} |h_{u,v}|^2 \tilde{r}_u^\alpha n_{\ell+1}^{-1} a_\ell^{\alpha/2} + N_0 + \Delta^2 \\ &\leq n_{\ell+1} 2^{-\alpha/2} a_\ell^{-\alpha/2} n_{\ell+1}^{-1} a_\ell^{\alpha/2} + N_0 + \Delta^2 \\ &= 2^{-\alpha/2} + N_0 + \Delta^2, \end{aligned}$$

where we have used (3.16). Hence

$$\begin{aligned} I(y_v; \hat{y}_v | \{h_{\tilde{u},\tilde{v}}\}) &= h(\hat{y}_v | \{h_{\tilde{u},\tilde{v}}\}) - h(\hat{y}_v | y_v, \{h_{\tilde{u},\tilde{v}}\}) \\ &\leq \log\big(2\pi e \mathbb{E}(|\hat{y}_v|^2)\big) - \log(2\pi e \Delta^2) \\ &\leq \log\big(2\pi e (2^{-\alpha/2} + N_0 + \Delta^2)\big) - \log(2\pi e \Delta^2) \\ &= \log\Big(1 + \frac{2^{-\alpha/2} + N_0}{\Delta^2}\Big). \end{aligned} \tag{3.17}$$

We now compute $I(x_u; \hat{x}_u | \{h_{\tilde{u},\tilde{v}}\})$. We have

$$\hat{x}_u = \|\boldsymbol{h}_u\| x_u + \sum_{\tilde{u} \in U(n_{\ell+1}) \setminus \{u\}} \frac{\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}}{\|\boldsymbol{h}_u\|} x_{\tilde{u}} + \frac{\boldsymbol{h}_u^\dagger}{\|\boldsymbol{h}_u\|}(\boldsymbol{z} + \tilde{\boldsymbol{z}}).$$

Conditioned on $\{\boldsymbol{h}_{\tilde{u}}\}_{\tilde{u}\in U(n_{\ell+1})}$,

$$\|\boldsymbol{h}_u\|x_u \sim \mathcal{N}_{\mathbb{C}}\big(0, \|\boldsymbol{h}_u\|^2 \tilde{r}_u^{\alpha} n_{\ell+1}^{-1} a_\ell^{\alpha/2}\big),$$

and

$$\mathbb{E}\Big(\Big|\sum_{\tilde{u}\in U(n_{\ell+1})\setminus\{u\}} \frac{\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}}{\|\boldsymbol{h}_u\|} x_{\tilde{u}} + \frac{\boldsymbol{h}_u^\dagger}{\|\boldsymbol{h}_u\|}(\boldsymbol{z}+\tilde{\boldsymbol{z}})\Big|^2 \Big| \{\boldsymbol{h}_{\tilde{u}}\}\Big) \\ = n_{\ell+1}^{-1} a_\ell^{\alpha/2} \sum_{\tilde{u}\in U(n_{\ell+1})\setminus\{u\}} \tilde{r}_{\tilde{u}}^{\alpha} \frac{|\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}|^2}{\|\boldsymbol{h}_u\|^2} + N_0 + \Delta^2,$$

where we have used the assumption that $\{z_v\}_{v\in V(n_{\ell+1})}$ are uncorrelated. Using (3.16), this is, in turn, upper bounded by

$$2^{3\alpha/2} \tilde{r}_u^{\alpha} n_{\ell+1}^{-2} a_\ell^{\alpha} \sum_{\tilde{u}\in U(n_{\ell+1})\setminus\{u\}} \tilde{r}_{\tilde{u}}^{\alpha} |\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}|^2 + N_0 + \Delta^2.$$

Similarly, we can lower bound the received signal power as

$$\mathbb{E}\big(\|\boldsymbol{h}_u\|^2 |x_u|^2\big) \geq 2^{-3\alpha/2}.$$

Since Gaussian noise is the worst additive noise under a power constraint [18], and applying Jensen's inequality to the convex function $\log(1+1/x)$, we obtain

$$\begin{aligned} &I(x_u; \hat{x}_u | \{h_{\tilde{u},\tilde{v}}\}) \\ &\quad \geq \mathbb{E}\left(\log\left(1 + \frac{2^{-3\alpha/2}}{2^{3\alpha/2} \tilde{r}_u^{\alpha} n_{\ell+1}^{-2} a_\ell^{\alpha} \sum_{\tilde{u}\in U(n_{\ell+1})\setminus\{u\}} \tilde{r}_{\tilde{u}}^{\alpha} |\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}|^2 + N_0 + \Delta^2}\right)\right) \\ &\quad \geq \log\left(1 + \frac{2^{-3\alpha/2}}{2^{3\alpha/2} \tilde{r}_u^{\alpha} n_{\ell+1}^{-2} a_\ell^{\alpha} \sum_{\tilde{u}\in U(n_{\ell+1})\setminus\{u\}} \tilde{r}_{\tilde{u}}^{\alpha} \mathbb{E}\big(|\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}|^2\big) + N_0 + \Delta^2}\right). \end{aligned} \quad (3.18)$$

We have for $u \neq \tilde{u}$,

$$\mathbb{E}\big(|\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}}|^2\big) = \mathbb{E}(\boldsymbol{h}_u^\dagger \boldsymbol{h}_{\tilde{u}} \boldsymbol{h}_{\tilde{u}}^\dagger \boldsymbol{h}_u)$$

$$= \sum_{v \in V(n_{\ell+1})} |h_{u,v}|^2 |h_{\tilde{u},v}|^2$$
$$= \sum_{v \in V(n_{\ell+1})} r_{u,v}^{-\alpha} r_{\tilde{u},v}^{-\alpha}, \tag{3.19}$$

and hence using (3.15)

$$\mathbb{E}\left(\tilde{r}_u^{\alpha} \sum_{\tilde{u} \in U(n_{\ell+1}) \setminus \{u\}} \tilde{r}_{\tilde{u}}^{\alpha} |\boldsymbol{h}_u^{\dagger} \boldsymbol{h}_{\tilde{u}}|^2 \right) = \sum_{\tilde{u} \in U(n_{\ell+1}) \setminus \{u\}} \sum_{v \in V(n_{\ell+1})} \tilde{r}_u^{\alpha} r_{u,v}^{-\alpha} \tilde{r}_{\tilde{u}}^{\alpha} r_{\tilde{u},v}^{-\alpha}$$
$$\leq 2^{-\alpha} n_{\ell+1}^2 a_{\ell}^{-\alpha}.$$

Therefore we can continue (3.18) as

$$I(x_u; \hat{x}_u | \{h_{\tilde{u},\tilde{v}}\}) \geq \log\left(1 + \frac{2^{-3\alpha/2}}{2^{\alpha/2} + N_0 + \Delta^2}\right) \triangleq K_4. \tag{3.20}$$

Using (3.17) and (3.20) in Lemma 3.8, and observing that we only communicate during a fraction

$$P_{\ell}(n) n_{\ell+1} a_{\ell}^{-\alpha/2} \leq 1$$

of time yields a per source node rate $\rho_{\ell}^{\mathsf{MAC}}(n)$ arbitrarily close to

$$K_4 P_{\ell}(n) n_{\ell+1} a_{\ell}^{-\alpha/2}$$

and a quantizer of rate arbitrarily close to

$$\log\left(1 + \frac{2^{-\alpha/2} + N_0}{\Delta^2}\right)$$

bits per observation at each relay node. Since by (3.20) the mutual information $I(x_u; \hat{x}_u | \{h_{\tilde{u},\tilde{v}}\})$ is at least K_4 for every $u \in U(n_{\ell+1})$ during the fraction of time we actually communicate, this implies that there are at most $1/K_4$ observations at each relay node per $n_{\ell+1}$ total message bits. Thus the number of bits per relay node

required to quantize the observations is at most

$$K_5 \triangleq \frac{1}{K_4} \log\left(1 + \frac{2^{-\alpha/2} + N_0}{\Delta^2}\right)$$

bits per $n_{\ell+1}$ total message bits sent by the source nodes. □

3.4.3 Broadcast Phase

At the end of the MAC phase, each node in the relay subsquare received a part of the message sent by each source node. In the BC phase, each node in the relay subsquare encodes these messages together for $n_{\ell+1}$ transmit antennas. The encoded message is then quantized and communicated to all the nodes in the relay subsquare. These nodes then send the quantized encoded message to the destination nodes $W(n_{\ell+1})$. Note that this again induces a uniform traffic pattern between the nodes in the relay subsquare, i.e., every node needs to transmit quantized encoded messages to every other node. While this traffic pattern does not correspond to a permutation traffic matrix it can be written as a sum of $n_{\ell+1}$ permutation traffic matrices. A $1/n_{\ell+1}$ fraction of the traffic within the relay subsquare is transmitted according to each of these permutation traffic matrices. This setup is depicted in Figure 3-3 in Section 3.2.1.

Assuming for the moment that we have a scheme to send the quantized encoded messages to the corresponding nodes in the relay subsquare, the traffic matrix $\widetilde{S}^{(1,1)}(n_{\ell+1})$ between $V(n_{\ell+1})$ and $W(n_{\ell+1})$ describes then a BC with one transmitter with $n_{\ell+1}$ antennas and $n_{\ell+1}$ receivers, each with one antenna. We call this the *BC induced by* $\widetilde{S}^{(1,1)}(n_{\ell+1})$ in the following.

Lemma 3.10. *For the BC induced by* $\widetilde{S}^{(1,1)}(n_{\ell+1})$ *with per-node average power constraint* $P_\ell(n) \leq n_{\ell+1}^{-1} a_\ell^{\alpha/2}$*, a rate of*

$$\rho_\ell^{\mathsf{BC}}(n) \geq K_6 P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}$$

is achievable per destination node, and the number of bits required to quantize the

observations is at most $K_7(\ell+1)\log(n)$ *bits at each relay node per* $n_{\ell+1}$ *total message bits*[5] *received by the destination nodes, for some constants* K_6 *and* K_7*.*

Proof. Consider a node $v \in V(n_{\ell+1})$ in the relay subsquare, say the first one. From the MAC phase, this node received the first part of the messages of each source node $u \in U(n_{\ell+1})$. We would like to jointly encode these message parts at the relay node using transmit beamforming, and then transmit the corresponding encoded signal using all the nodes in the relay subsquare. However, this cannot be done directly because at the encoding time, the future channel state at transmission time is unknown.

We circumvent this problem by reordering the signals to be transmitted at the relay nodes as follows. Let

$$\{\hat{\theta}_{v,w}\}_{v\in V(n_{\ell+1}),w\in W(n_{\ell+1})} \in \{0,\pi/2,\pi,3\pi/2\}^{n_{\ell+1}^2}.$$

be a "quantized" channel state. The part of the messages at node v in the relay subsquare is encoded for $n_{\ell+1}$ transmit nodes with an assumed channel gain of

$$\hat{h}_{v,w}[t] = r_{v,w}^{-\alpha/2}\exp(\sqrt{-1}\hat{\theta}_{v,w}[t]),$$

where the $\{\hat{\theta}_{v,w}[t]\}_{v,w,t}$ are cycled as a function of t through all possible values in $\{0,\pi/2,\pi,3\pi/2\}^{n_{\ell+1}^2}$. The components of the encoded messages are then quantized and each component sent to the corresponding node in the relay subsquare. Once all nodes in the relay subsquare have received the encoded message, they send in each time slot a sample of the encoded messages corresponding to the quantized channel state closest (in Euclidean distance) to the actual channel realization in that time slot. By ergodicity of $\{\theta_{u,v}[t]\}_t$, each quantized channel state is used approximately the same number of times. More precisely, as the message length grows to infinity, we can send samples of the encoded message parts a $1/(1+\eta)$ fraction of time with probability approaching 1 for any $\eta > 0$. Since we have no constraint on the encoding

[5]Total message bits refers to the sum of all message bits received by the $n_{\ell+1}$ destination nodes.

delay in our setup, we can choose η arbitrarily small, and given that we are only interested in scaling laws, we will ignore this term in the following to simplify notation. Note that the destination nodes can reorder the received samples since we assume full CSI. In the following, we let $\{\hat{\theta}_{v,w}\}_{v,w}$ be the random quantized channel state induced by $\{\theta_{v,w}\}_{v,w}$ through the above procedure. Denote by $\{\hat{h}_{v,w}\}_{v,w}$ the corresponding channel gains.

As in the MAC phase, the nodes in the relay subsquare send signals at a power (essentially) $n_{\ell+1}^{-1} a_\ell^{\alpha/2}$ a fraction $P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2} \leq 1$ of time and are silent for the remaining time. To create interference at uniform power, this is done in the same randomized manner as in the MAC phase. Generate independently for each region $A(a_\ell)$ a Bernoulli process $\{B[t]\}_{t\in\mathbb{N}}$ with parameter $P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}/(1+\eta)$ for some small $\eta > 0$. The nodes in $A(a_\ell)$ are active whenever $B[t] = 1$ and remain silent otherwise. As before, we ignore the additional $1/(1+\eta)$ term. Again we only need to consider the fraction of time during which $B[t] = 1$.

Consider the message part at a relay node for destination node $w \in W(n_{\ell+1})$. We encode this part independently; call $\tilde{x}_w$ the encoded message part. The relay node then performs transmit beamforming to construct the encoded message for all its destination nodes, i.e.,

$$\boldsymbol{x} = \sum_{w \in W(n_{\ell+1})} \frac{\hat{\boldsymbol{h}}_w^\dagger}{\|\boldsymbol{h}_w\|} \tilde{x}_w,$$

where row vector $\boldsymbol{h}_w = \{h_{v,w}\}_{v\in V(n_{\ell+1})}$ contains the channel gains to node w, and where we have used $|\hat{h}_{v,w}| = |h_{v,w}|$. The relay node then quantizes the vector of encoded messages componentwise and forwards the quantized version $\hat{\boldsymbol{x}}$ to the other nodes in the relay subsquare. These nodes then send $\hat{\boldsymbol{x}}$ over the channel to the destination nodes. The received signal at destination node w is thus

$$y_w = \boldsymbol{h}_w \hat{\boldsymbol{x}} + z_w.$$

With this, we have the setup considered in Lemma 3.8 (with different variable names). The first "channel" in Lemma 3.8 (see Figure 3-5) will correspond to the

transmit beamforming used at the relay subsquare. The second channel in Lemma 3.8 will now correspond to the wireless channel between the relay subsquare $V(n_{\ell+1})$ and a destination node w. To apply Lemma 3.8, we need to find a distribution for $\tilde{x}_w$ and for $\hat{x}_v|x_v$. We also need to guarantee that $\hat{x}_v$ satisfies the power constraint at each node v in the relay subsquare. For each $w \in W(n_{\ell+1})$, let

$$\tilde{x}_w \sim \mathcal{N}_{\mathbb{C}}(0, Kn_{\ell+1}^{-1}a_\ell^{\alpha/2})$$

(for some K to be chosen later) independent of $\tilde{x}_{\tilde{w}}$ for $w \neq \tilde{w}$, and let $\hat{x}_v = x_v + \tilde{z}_v$ for $\tilde{z}_v \sim \mathcal{N}_{\mathbb{C}}(0, \Delta^2)$ independent of $\boldsymbol{x}$ and for some $\Delta^2 > 0$. We then have

$$y_w = \frac{\boldsymbol{h}_w \hat{\boldsymbol{h}}_w^\dagger}{\|\boldsymbol{h}_w\|}\tilde{x}_w + \sum_{\tilde{w}\in W(n_{\ell+1})\setminus\{w\}} \frac{\boldsymbol{h}_w \hat{\boldsymbol{h}}_{\tilde{w}}^\dagger}{\|\boldsymbol{h}_{\tilde{w}}\|}\tilde{x}_{\tilde{w}} + \boldsymbol{h}_w\tilde{\boldsymbol{z}} + z_w.$$

We proceed by computing $I(x_v; \hat{x}_v|\{h_{\tilde{u},\tilde{v}}\})$ and $I(\tilde{x}_w; y_w|\{h_{\tilde{u},\tilde{v}}\})$ as required in Lemma 3.8 (the conditioning on $\{h_{\tilde{u},\tilde{v}}\}$ is again due to availability of full CSI). Note first that by construction of $\widetilde{S}^{(1,1)}(n_{\ell+1})$, we have for any $w \in W(n_{\ell+1})$

$$2 \min_{v\in V(n_{\ell+1})} r_{v,w} \geq \max_{v\in V(n_{\ell+1})} r_{v,w},$$

and therefore

$$\frac{|h_{v,w}|^2}{\|\boldsymbol{h}_w\|^2} \leq \frac{\big(\min_{v\in V(n_{\ell+1})} r_{v,w}\big)^{-\alpha}}{n_{\ell+1}\big(\max_{v\in V(n_{\ell+1})} r_{v,w}\big)^{-\alpha}} \leq \frac{2^\alpha}{n_{\ell+1}}. \tag{3.21}$$

We start by computing $I(x_v; \hat{x}_v|\{h_{\tilde{u},\tilde{v}}\})$. $\hat{x}_v$ has mean zero and variance

$$\begin{aligned}
\mathbb{E}\big(|\hat{x}_v|^2\big) &= \sum_{w\in W(n_{\ell+1})} \frac{|h_{v,w}|^2}{\|\boldsymbol{h}_w\|^2} Kn_{\ell+1}^{-1}a_\ell^{\alpha/2} + \Delta^2 \\
&\leq n_{\ell+1}\frac{2^\alpha}{n_{\ell+1}} Kn_{\ell+1}^{-1}a_\ell^{\alpha/2} + \Delta^2 \\
&\leq n_{\ell+1}^{-1}a_\ell^{\alpha/2},
\end{aligned} \tag{3.22}$$

for

$$K \triangleq 2^{-\alpha}(1-\Delta^2),$$

which is positive for $\Delta^2 < 1$, and where we have used (3.21) and that

$$n_{\ell+1}^{-1} a_\ell^{\alpha/2} \geq 2^{\ell+1}\gamma(n) \geq 1$$

by (3.12). Equation (3.22) shows that $\hat{x}_v$ satisfies the power constraint of node v in the relay subsquare $V(n_{\ell+1})$. Moreover, we obtain

$$\begin{aligned} I(x_v; \hat{x}_v | \{h_{\tilde{u},\tilde{v}}\}) &= h(\hat{x}_v | \{h_{\tilde{u},\tilde{v}}\}) - h(\hat{x}_v | x_v, \{h_{\tilde{u},\tilde{v}}\}) \\ &\leq \log\Big(2\pi e \mathbb{E}\big(|\hat{x}_v|^2\big)\Big) - \log(2\pi e \Delta^2) \\ &\leq \log\left(\frac{n_{\ell+1}^{-1} a_\ell^{\alpha/2}}{\Delta^2}\right). \end{aligned} \tag{3.23}$$

It remains to compute $I(\tilde{x}_w; y_w | \{h_{\tilde{u},\tilde{v}}\})$. Note that the encoding procedure guarantees that

$$\cos(\pi/4)^2 \|\boldsymbol{h}_w\|^4 \leq |\boldsymbol{h}_w \hat{\boldsymbol{h}}_w^\dagger|^2 \leq \|\boldsymbol{h}_w\|^4.$$

Moreover, for $w \neq \tilde{w}$,

$$\begin{aligned} \mathbb{E}\big(|\boldsymbol{h}_w \hat{\boldsymbol{h}}_{\tilde{w}}^\dagger|^2\big) &= \mathbb{E}(\boldsymbol{h}_w \hat{\boldsymbol{h}}_{\tilde{w}}^\dagger \hat{\boldsymbol{h}}_{\tilde{w}} \boldsymbol{h}_w^\dagger) \\ &= \sum_{v \in V(n_{\ell+1})} \mathbb{E}\big(|h_{vw}|^2 |\hat{h}_{v\tilde{w}}|^2\big) \\ &= \sum_{v \in V(n_{\ell+1})} \mathbb{E}\big(|h_{vw}|^2 |h_{v\tilde{w}}|^2\big) \\ &= \mathbb{E}\big(|\boldsymbol{h}_w \boldsymbol{h}_{\tilde{w}}^\dagger|^2\big). \end{aligned}$$

From this, we get by a similar argument as in Lemma 3.9 that

$$I(\tilde{x}_w; y_w | \{h_{\tilde{u},\tilde{v}}\}) \geq K_6. \tag{3.24}$$

Using (3.23) and (3.24) in Lemma 3.8, and observing that we only communicate during a fraction

$$P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}$$

of time, yields a per destination node rate $\rho_\ell^{\mathsf{BC}}(n)$ arbitrarily close to

$$K_6 P_\ell(n) n_{\ell+1} a_\ell^{-\alpha/2}$$

bits per channel use and a quantizer rate arbitrarily close to

$$\log\Big(\frac{n_{\ell+1}^{-1} a_\ell^{\alpha/2}}{\Delta^2}\Big)$$

bits per encoded sample. Since by (3.24) mutual information $I(\tilde{x}_w; y_w|\{h_{\tilde{u},\tilde{v}}\})$ is at least K_6 for every $w \in W(n_{\ell+1})$ during the fraction of time we actually communicate, this implies that there are at most $1/K_6$ encoded message samples for each relay node per $n_{\ell+1}$ total message bits received by the destination nodes $W(n_{\ell+1})$. Thus the number of bits required at each relay node to quantize the encoded message samples is at most

$$\begin{aligned}\frac{1}{K_6}\log\Big(\frac{n_{\ell+1}^{-1} a_\ell^{\alpha/2}}{\Delta^2}\Big) &= \frac{1}{K_6}\log\Big(\frac{1}{\Delta^2} 2^{\ell+1}\gamma^{1+\ell(1-\alpha/2)}(n) n^{\alpha/2-1}\Big)\\ &\le \frac{1}{K_6}\log\Big(\frac{1}{\Delta^2} 2^{\ell+1} n^{\alpha/2}\Big)\\ &\le K_7(\ell+1)\log(n)\end{aligned}$$

bits per $n_{\ell+1}$ total message bits received by the destination nodes, and where we have used $\gamma(n) \le n$ by (3.12). □

3.5 Proof of Theorem 3.1

The proof of Theorem 3.1 is split into two parts. In Section 3.5.1 we prove the theorem for fast fading, and in Section 3.5.2 for slow fading.

3.5.1 Fast Fading

In this section, we prove Theorem 3.1 under fast fading, i.e., $\{\theta_{u,v}[t]\}_t$ is stationary and ergodic in t. We first prove that the assumptions on the power constraint and

the interference made in Section 3.4 (see Lemmas 3.9 and 3.10) during the analysis of one level of the hierarchical relaying scheme are valid. We then use the results proved there to analyze the behavior of the entire hierarchy, yielding a lower bound on the per-node rate achievable with hierarchical relaying.

We first argue that the constraint $P_\ell(n) \le n_{\ell+1}^{-1} a_\ell^{\alpha/2}$ needed in Lemmas 3.9 and 3.10 is satisfied. Consider the hierarchical relaying scheme as described in Section 3.2, and fix a level ℓ, $0 \le \ell < L = L(n)$, in this hierarchy. At level ℓ, we have a square of area $a_\ell = n/\gamma^\ell(n)$ with $n_\ell = n/2^\ell\gamma^\ell(n)$ source-destination pairs. Since we are time sharing between $K_2 2^{-\ell}\gamma(n)$ relay subsquares at this level, we have an average power constraint of

$$P_\ell(n) \triangleq K_2 2^{-\ell}\gamma(n)$$

during the time any particular relay subsquare is active. Since $\alpha > 2$ and since $n\gamma^{-L(n)}(n) \to \infty$ as $n \to \infty$, we have, for n large enough (independent of ℓ), that

$$\begin{aligned} P_\ell(n) &= K_2 2^{-\ell}\gamma(n) \\ &\le 2^{-\ell}\gamma(n)\Big(\frac{n}{\gamma^{L(n)}(n)}\Big)^{\alpha/2-1} \\ &\le 2^{\ell+1}\gamma(n)\Big(\frac{n}{\gamma^{\ell}(n)}\Big)^{\alpha/2-1} \\ &= n_{\ell+1}^{-1} a_\ell^{\alpha/2}. \end{aligned}$$

Therefore the power constraints in Lemmas 3.9 and 3.10 are satisfied.

We continue by analyzing the interference caused by spatial re-use. Recall that the MAC and BC phases at level ℓ induce permutation traffic within the dense subsquares at level $\ell+1$. The permutation traffic within those dense subsquares at level $\ell+1$ is transmitted in parallel with spatial re-use. We now describe in detail how this spatial re-use is performed. Partition the subsquares of area $a_{\ell+1}$ (i.e., at level $\ell+1$) into four subsets such that in each subset all subsquares are at distance at least $\sqrt{a_{\ell+1}}$ from each other. The traffic that the MAC and BC phases at level ℓ induce in each of the relay subsquares at level $\ell+1$ is transmitted simultaneously within all relay subsquares in the same subset. Consider now one such subset. We show that at any

relay subsquare the interference from other relay subsquares in the same subset is stationary and ergodic within each phase, additive (i.e., independent of the signals and channel gains in this relay subsquare), and of bounded power $N_0 - 1$ independent of n.

We first argue that the interference is stationary and ergodic within each phase. Note first that on any level $\ell + 1$ in the hierarchy, all relay subsquares are either simultaneously in the MAC phase or simultaneously in the BC phase. Furthermore, all relay subsquares are also synchronized for transmissions within each of these phases (recall that the induced traffic in level $\ell + 1$ is uniform and is sent sequentially as permutation traffic). Hence it suffices to show that the interference generated by either the MAC or BC induced by some permutation traffic matrix is stationary and ergodic. Since all codebooks for either of these cases are generated as i.i.d. Gaussian multiplied by a Bernoulli process, and in the BC phase beamformed for stationary and ergodic fading, this is indeed the case.

The additivity of the interference follows easily for the MAC phase since codebooks are generated independently of the channel realization in this case. Moreover, since the channel gains are independent from each other and all codebooks are generated as independent zero mean processes, the interference in the MAC phase is also uncorrelated (over space) within each relay subsquare. For the BC phase, the codebook depends only on the channel gains within each relay subsquare at level $\ell + 1$. Since the channel gains within relay subsquares are independent of the channel gains between relay subsquares, this interference is additive as well.

We now bound the interference power. Note that by the randomized time-sharing construction within the MAC and BC phases (see Lemmas 3.9 and 3.10), in each relay subsquare at most $n_{\ell+1}$ nodes transmit at an average power of 1. In the MAC phase, all nodes use independently generated codebooks with power at most 1, and thus the received interference power from another relay subsquare at distance $i\sqrt{a_{\ell+1}}$ is at most

$$n_{\ell+1} i^{-\alpha} a_{\ell+1}^{-\alpha/2} = i^{-\alpha} 2^{-(\ell+1)} \Big(\frac{n}{\gamma^{\ell+1}(n)} \Big)^{1-\alpha/2} \leq i^{-\alpha},$$

by (3.12). In the BC phase, the nodes in each active relay subsquare use beamforming to transmit to nodes within their own subsquare. Since the channel gains within a relay subsquare are independent of the channel gains between relay subsquares, the same calculation as in (3.19) shows that we can upper bound the received interference power from another relay subsquare at distance $i\sqrt{a_{\ell+1}}$ by

$$n_{\ell+1} i^{-\alpha} a_{\ell+1}^{-\alpha/2} \leq i^{-\alpha},$$

in the BC phase as well.

Now, by the way in which we perform spatial re-use, every active relay subsquare has at most $8i$ active relay subsquares at distance at least $i\sqrt{a_{\ell+1}}$. Hence the total interference power received at an active relay subsquare is at most

$$\sum_{i=1}^{\infty} 8i 2^{\alpha} i^{-\alpha} \triangleq N_0 - 1 < \infty$$

since $\alpha > 2$. With this, we have shown that the interference term has the properties required in Lemmas 3.9 and 3.10.

We now apply those two lemmas to obtain a lower bound on the rate achievable with hierarchical relaying. Call $\tau_\ell(n)$ the number of channel uses to transmit one bit from each of n_ℓ source nodes to the corresponding destination nodes at level ℓ. Lemma 3.7 states that for n large enough (independent of ℓ), we relay over each dense subsquare at most $K_3 2^\ell$ times. Combining this with Lemma 3.9, we see that to transmit one bit from each source to its destination at this level we need at most

$$4K_3 2^\ell K_2 2^{-\ell} \gamma(n) \frac{1}{K_4 P_\ell(n)} n_{\ell+1}^{-1} a_\ell^{\alpha/2} = \frac{K_3 2^{2\ell+3}}{K_4} n^{\alpha/2-1} \gamma^{1+\ell(1-\alpha/2)}(n)$$

channel uses for the MAC phase. Here, the factor 4 accounts for the spatial re-use, $K_3 2^\ell$ accounts for relaying over the same relay subsquares multiple times, $K_2 2^{-\ell}\gamma(n)$ accounts for time sharing between the relay subsquares, and the last term accounts for the time required to communicate over the MAC. Similarly, combining Lemmas 3.7

and 3.10, we need at most

$$\frac{K_3 2^{2\ell+3}}{K_6} n^{\alpha/2-1} \gamma^{1+\ell(1-\alpha/2)}(n)$$

channel uses for the BC phase. Moreover, at level $\ell+1$ in the hierarchy this induces a per-node traffic demand of at most K_5 bits from the MAC phase, and at most $K_7(\ell+1)\log(n)$ from the BC phase. Thus we obtain the following recursion

$$\begin{aligned}
\tau_\ell(n) &\leq 8K_3\Big(\frac{1}{K_4}+\frac{1}{K_6}\Big)n^{\alpha/2-1}\gamma(n)\big(4\gamma^{1-\alpha/2}(n)\big)^\ell + (K_5 + K_7(\ell+1)\log(n))\tau_{\ell+1}(n) \\
&\leq \widetilde{K} n^{\alpha/2-1}\gamma(n)4^\ell + K(\ell+1)\log(n)\tau_{\ell+1}(n) \\
&\leq \widetilde{K} n^{\alpha/2-1}\gamma(n)4^{L(n)} + KL(n)\log(n)\tau_{\ell+1}(n) \qquad (3.25)
\end{aligned}$$

for positive constants $K, \widetilde{K}$ independent of n and ℓ.

At scale a_L, we have n_L nodes and source-destination pairs. Time sharing between all source-destination pairs, we have (during the time we communicate for each node) an average power constraint of n_L. Since at this level we communicate over a distance of at most $2a_L^{1/2}$, we have

$$\tau_L(n) \leq n_L \log^{-1}\left(1+\frac{n_L}{2^\alpha N_0 a_L^{\alpha/2}}\right). \qquad (3.26)$$

Since

$$n_L a_L^{-\alpha/2} \leq n_L a_L^{-1} = 2^{-L(n)} \to 0$$

as $n \to \infty$, we can upper bound (3.26) as

$$\begin{aligned}
\tau_L(n) &\leq K' a_L^{\alpha/2} \\
&= K' n^{\alpha/2}\gamma^{-L(n)\alpha/2}(n) \\
&\leq K' n^{\alpha/2}\gamma^{-L(n)}(n) \qquad (3.27)
\end{aligned}$$

for some constant K'.

Now, using the recursion (3.25) $L(n)$ times, and combining with (3.27), we obtain

$$\begin{aligned}
\tau_0(n) &\leq \widetilde{K} n^{\alpha/2-1}\gamma(n)4^{L(n)} + KL(n)\log(n)\tau_1(n) \\
&\leq \dots \\
&\leq \widetilde{K} n^{\alpha/2-1}\gamma(n)4^{L(n)}\Bigg(\sum_{\ell=0}^{L(n)-1}\big(KL(n)\log(n)\big)^{\ell}\Bigg) + \big(KL(n)\log(n)\big)^{L(n)}\tau_L(n) \\
&\leq n^{\alpha/2-1}\big(KL(n)\log(n)\big)^{L(n)}\Big(\widetilde{K}4^{L(n)}\gamma(n) + K' n\gamma^{-L(n)}(n)\Big). \qquad (3.28)
\end{aligned}$$

Using the definition of $\gamma(n)$ and $L(n)$ in (3.11), we have for n large enough

$$\begin{aligned}
\big(KL(n)\log(n)\big)^{L(n)} &\leq n^{2\log^{-1/2-\delta}(n)\log\log(n)}, \\
4^{L(n)}\gamma(n) &\leq n^{2\log^{-1/2-\delta}(n)+\log^{\delta-1/2}(n)}, \\
n\gamma^{-L(n)}(n) &\leq n^{\log^{\delta-1/2}(n)}.
\end{aligned}$$

Since $\delta > 0$, the $n^{\log^{\delta-1/2}(n)}$ term dominates in (3.28), and we obtain

$$\tau_0(n) \leq \tilde{b}(n) n^{\alpha/2-1},$$

where

$$\tilde{b}(n) \leq n^{O(\log^{\delta-1/2}(n))},$$

as $n \to \infty$. Therefore

$$\rho^*(n) \geq \rho^{\mathsf{HR}}(n) = 1/\tau_0(n) \geq b(n) n^{1-\alpha/2},$$

with

$$b(n) \geq n^{-O(\log^{\delta-1/2}(n))},$$

concluding the proof for the fast fading case.

3.5.2 Slow Fading

In this section, we prove Theorem 3.1 under slow fading, i.e., $\{\theta_{u,v}[t]\}_t$ is constant as a function of t. We sketch the necessary modifications for the scheme described in Section 3.2 to achieve a per-node rate of at least $b(n)n^{1-\alpha/2}$ in the slow fading case.

Consider level ℓ, $0 \leq \ell < L(n)$, in the hierarchy. Instead of relaying the message of a source-destination pair over one relay subsquare as in the scheme described in Section 3.2, we relay the message over many dense subsquares that are at least at distance $\sqrt{2a_{\ell+1}}$ from both the source and the destination nodes. We time share between the different relays. The idea here is that the wireless channel between any node and its relay subsquare might be in a bad state due to the slow fading, making communication over this relay subsquare impossible. Averaged over many relay subsquares, however, we get essentially the same performance as in the fast fading case.

We first state a (somewhat weaker) version of Lemma 3.7, appropriate for this setup. Consider again the collection of schedules $\mathcal{S}(n_\ell)$ and $\widetilde{\mathcal{S}}(n_\ell)$ satisfying the conditions that no relay subsquare is selected by more than $n_{\ell+1}$ source-destination pairs and that all sources and destinations are at least at distance $\sqrt{2a_{\ell+1}}$ from their relay subsquare (see Section 3.4.1 for the formal definition). The next lemma shows that for each source-destination pair, we can find $K_2 2^{-\ell-1}\gamma(n)$ distinct relay subsquares satisfying the above conditions (the requirement that these relay subsquares are distinct is expressed by the orthogonality condition of the schedules in Lemma 3.11 below).

Lemma 3.11. *For every n large enough (independent of ℓ) and every permutation traffic matrix $\lambda^{\mathsf{UC}}(n_\ell) \in \{0,1\}^{n_\ell \times n_\ell}$ there are schedules $\{S^{(i)}(n_\ell)\}_{i=1}^{K_2 2^{-\ell}\gamma^2(n)} \subset \mathcal{S}(n_\ell)$, $\{\widetilde{S}^{(i)}(n_\ell)\}_{i=1}^{K_2 2^{-\ell}\gamma^2(n)} \subset \widetilde{\mathcal{S}}(n_\ell)$ satisfying*

$$\lambda^{\mathsf{UC}}(n_\ell) = \frac{1}{K_2 2^{-\ell-1}\gamma(n)} \sum_{i=1}^{K_2 2^{-\ell}\gamma^2(n)} S^{(i)}(n_\ell)\widetilde{S}^{(i)}(n_\ell),$$

where $\{S^{(i)}(n_\ell)\}_i$, $\{\widetilde{S}^{(i)}(n_\ell)\}_i$ are collections of orthogonal matrices in the sense that

for $i \neq i'$,

$$\sum_{u,k} s_{u,k}^{(i)} s_{u,k}^{(i')} = 0,$$
$$\sum_{k,u} \tilde{s}_{k,u}^{(i)} \tilde{s}_{k,u}^{(i')} = 0. \tag{3.29}$$

Proof. The proof is similar to that of Lemma 3.7. In order to construct $\{S^{(i)}(n_\ell)\}$ and $\{\widetilde{S}^{(i)}(n_\ell)\}$, consider the sequential pass over all n source-destination pairs (assume n is large enough for Lemma 3.7 to hold). As before, for each source-destination pair, there are $K_2 2^{-\ell-1}\gamma(n)$ dense relay subsquares that are at distance at least $\sqrt{2a_{\ell+1}}$. Each pair chooses all of these $K_2 2^{-\ell-1}\gamma(n)$ subsquares instead of just one as before. Stop one round of this procedure as soon as any of the relay subsquares is chosen by $n_{\ell+1}$ pairs. Since by the end of one round at least one relay subsquare is matched by $n_{\ell+1}$ source-destination pairs, there are at most $n_\ell/n_{\ell+1} = 2\gamma(n)$ such rounds.

Consider now the result of one such round. We construct $K_2 2^{-\ell-1}\gamma(n)$ matrices $S^{(i)}(n_\ell)$ and $\widetilde{S}^{(i)}(n_\ell)$, with the i-th pair of matrices describing communication over the i-th relay subsquares chosen by source-destination pairs matched in this round. Thus, this process produces a total of $2\gamma(n)K_2 2^{-\ell-1}\gamma(n) = K_2 2^{-\ell}\gamma^2(n)$ such matrices. The orthogonality property follows since each source-destination pair relays over the same relay subsquare at most once. □

Having decomposed the scaled traffic matrix $K_2 2^{-\ell-1}\gamma(n)\lambda^{\mathsf{UC}}(n)$ into $K_2 2^{-\ell}\gamma^2(n)$ matrices, each source-destination pair tries to relay over $K_2 2^{-\ell-1}\gamma(n)$ dense subsquares. We time share between these relay subsquares. Since each source-destination pair relays only a $(K_2 2^{-\ell-1}\gamma(n))^{-1}$ fraction of traffic over any of its relay subsquares, the loss due to this time sharing is now

$$\frac{K_2 2^{-\ell}\gamma^2(n)}{K_2 2^{-\ell-1}\gamma(n)} = 2\gamma(n)$$

as opposed to $K_3 2^\ell$ in Lemma 3.7. In other words, the loss is at most a factor $2\gamma(n)$ more than in Lemma 3.7. Using the definition of $\gamma(n)$ in (3.11), we have

$$\gamma(n) \leq n^{-\log^{\delta-1/2}(n)} \leq b^{-1}(n).$$

In other words, this additional loss is small.

Consider now a specific relay subsquare. If a source-destination pair can communicate over this relay subsquare at a rate at least 1/64-th of the rate achievable in the fast fading case (given by Lemmas 3.9 and 3.10), it sends information over this relay. Otherwise it stays silent during the period of time it is assigned this relay. We now show that, with probability $1 - o(1)$ as $n \to \infty$, for every source-destination pair on every level of the hierarchy at least one quarter of its relay subsquares can support this rate. As we only communicate over a quarter of the relay subsquares, this implies that we can achieve at least 1/256-th of the per-node rate for the fast fading case (see Section 3.5.1), i.e., that $b(n)n^{1-\alpha/2}$ is achievable with probability $1 - o(1)$ as $n \to \infty$.

Assume we have for each source-destination pair (u, w) picked $K_2 2^{-\ell-1}\gamma(n)$ dense subsquares over which it can relay; call those relay subsquares $\{A_{u,w,k}\}_{k=1}^{K_2 2^{-\ell-1}\gamma(n)}$. Consider the event $B_{u,w,k}$ that source node u can communicate at the desired rate to destination node w over relay subsquares $A_{u,w,k}$ (assuming, as before, that we can solve the communication problem within this subsquare).

Let $\{B_{u,w,k}^{(i)}\}_{i=1}^{4}$ be the events that the interference due to matched filtering in the MAC phase, the interference from spatial re-use in the MAC phase, the interference due to beamforming in the BC phase, and the interference from spatial re-use in the BC phase, are less than 8 times the one for fast fading, respectively. From the proof of Lemmas 3.9, 3.10, and of Theorem 3.1 for the fast fading case in Section3.5.1, we see that

$$\bigcap_{i=1}^{4} B_{u,w,k}^{(i)} \subset B_{u,w,k}.$$

Due to spatial re-use, multiple relay subsquares will be active in parallel. Let $\widetilde{H}$ denote the set of channel gains between active relay subsquares. Using essentially the same arguments as for the fast fading case (see Lemmas 3.9, 3.10, and Section 3.5.1) and from Markov's inequality, we have $\mathbb{P}(B_{u,w,k}^{(i)}|\widetilde{H}) \geq 7/8$ for all $i \in \{1, \ldots, 4\}$ and hence $\mathbb{P}(B_{u,w,k}|\widetilde{H}) \geq 1/2$.

We now argue that the events

$$\left\{ \cap_{i=1}^{4} B_{u,w,k}^{(i)} \right\}_{k=1}^{K_2 2^{-\ell-1}\gamma(n)} \tag{3.30}$$

are independent conditioned on $\widetilde{H}$, by showing that these events depend on disjoint sets of channel gains and codebooks. Assuming the codebooks are generated new for each communication round, then they are all independent. Thus we only have to consider the dependence on the channel gains. Let U_k and W_k be the source and destination nodes communicating over relay subsquare $A_{u,w,k}$ in round k, and let V_k be the nodes in $A_{u,w,k}$. Let $\widetilde{U}_k$, $\widetilde{W}_k$ be the source and destination nodes that are communicating at the same time as (u, w) due to spatial re-use. Let $\widetilde{V}_k$ be the relay nodes of $\widetilde{U}_k$ and $\widetilde{W}_k$. Now, $B_{u,w,k}^{(1)}$ and $B_{u,w,k}^{(2)}$ depend (for fixed $\widetilde{H}$) on the channel gains between U_k and V_k. $B_{u,w,k}^{(3)}$ depends on the channel gains between V_k and W_k. $B_{u,w,k}^{(4)}$ depends (again for fixed $\widetilde{H}$) on the channel gains between $\widetilde{V}_k$ and $\widetilde{W}_k$. Since these sets are disjoint for different k by the orthogonality of the schedules (see (3.29)), conditional independence of the events in (3.30) follows.

To summarize, conditioned on the channel gains $\widetilde{H}$ between active relay subsquares, the random variables $\{\mathbb{1}_{B_{u,w,k}}\}_k$ are independent and have expected value $\mathbb{E}(\mathbb{1}_{B_{u,w,k}}|\widetilde{H}) \geq 1/2$. The sum

$$\sum_{k=1}^{K_2 2^{-\ell-1}\gamma(n)} \mathbb{1}_{B_{u,w,k}}$$

is the number of relay subsquares over which the source-destination pair (u, w) successfully relays traffic. We now show that with high probability at least one quarter of these relay subsquares allow successful transmission. Indeed, by the Chernoff bound,

$$\begin{aligned}
\mathbb{P}\Big(\textstyle\sum_k \mathbb{1}_{B_{u,w,k}} < K_2 2^{-\ell-3}\gamma(n)\Big|\widetilde{H}\Big) &\leq \mathbb{P}\Big(\textstyle\sum_k \mathbb{1}_{B_{u,w,k}} < K_2 2^{-\ell-2}\gamma(n)\mathbb{P}(B_{u,w,k}|\widetilde{H})\Big|\widetilde{H}\Big) \\
&\leq \exp\big(-2K2^{-\ell}\gamma(n)\mathbb{P}(B_{u,w,k}|\widetilde{H})\big) \\
&\leq \exp\big(-K2^{-\ell}\gamma(n)\big)
\end{aligned}$$

for some constant $K > 0$. Since the right-hand side is the same for all $\widetilde{H}$, this implies

$$\mathbb{P}\Big(\textstyle\sum_k \mathbb{1}_{B_{u,w,k}} < K_2 2^{-\ell-3}\gamma(n)\Big) \leq \exp\big(-K2^{-\ell}\gamma(n)\big).$$

In each of the $L(n)$ levels of the hierarchy there are at most n^2 source-destination pairs, and hence by the union bound with probability at least

$$1 - L(n)n^2 \exp\big(-K2^{-L(n)}\gamma(n)\big),$$

for every source-destination pair on every level of the hierarchy at least one quarter of its relay subsquares can support the desired rate. By the choices of $\gamma(n)$ and $L(n)$ in (3.11), this probability is at least

$$\begin{aligned}
1 - L(n)n^2 \exp\big(-K2^{-L(n)}\gamma(n)\big) &\geq 1 - n^3 \exp\Big(-K2^{-L(n)}2^{\log(n)/2L(n)}\Big) \\
&\geq 1 - \exp\Big(\widetilde{K}2^{\log\log(n)} - K2^{\frac{1}{2}\log^{1/2+\delta}(n) - \log^{1/2-\delta}(n)}\Big) \\
&\geq 1 - \exp\Big(-2^{\Omega(\log^{1/2+\delta}(n))}\Big) \\
&\geq 1 - o(1)
\end{aligned}$$

as $n \to \infty$, and for some constant $\widetilde{K}$. This proves that the same order rate as in the fast fading case can be achieved with high probability for all levels $0 \leq \ell < L(n)$.

It remains to argue that the same holds for level $\ell = L(n)$. Note that since we assume phase fading only, the received signal power is only a function of distance and not of the fading realization. Since at level $L(n)$ we use simple time sharing, this implies that we can always achieve the same rate at level $L(n)$ as in the fast fading case.

Hence with probability $1 - o(1)$ as $n \to \infty$, we achieve the same order rate at each level $0 \leq \ell \leq L(n)$ as for fast fading, proving Theorem 3.1 for the slow fading case.

3.6 Proof of Theorem 3.2

Here, we provide a generalization and sharpening of the converse in [38]. Most of the arguments follow [38, Theorem 5.2]. We start by proving a lemma upper bounding the MIMO capacity.

Consider two subsets $S_1, S_2 \subset V(n)$ such that $S_1 \cap S_2 = \emptyset$. Assume we allow the nodes within S_1 and S_2 to cooperate without any restriction. The maximum achievable sum rate between the nodes in S_1 and S_2 is given by the MIMO capacity $C(S_1, S_2)$ between them. The next lemma upper bounds $C(S_1, S_2)$ in terms of the node distances between the two sets and the *normalized channel gains*

$$\tilde{h}_{u,v} \triangleq \frac{h_{u,v}}{\sqrt{\sum_{\tilde{v} \in S_2} r_{u,\tilde{v}}^{-\alpha}}}. \tag{3.31}$$

Lemma 3.12. *Under either fast or slow fading, for every $\alpha > 2$, for any node placement $V(n)$, and any $S_1, S_2 \subset V(n)$ with $S_1 \cap S_2 = \emptyset$, we have*

$$C(S_1, S_2) \leq 4 \Bigg(\max \Big\{ 1, \max_{v \in S_2} \sum_{u \in S_1} |\tilde{h}_{u,v}|^2 \Big\} \Bigg) \sum_{u \in S_1} \sum_{v \in S_2} r_{u,v}^{-\alpha}.$$

Proof. Let

$$\boldsymbol{H} \triangleq \{h_{u,v}\}_{u \in S_1, v \in S_2},$$
$$\widetilde{\boldsymbol{H}} \triangleq \{\tilde{h}_{u,v}\}_{u \in S_1, v \in S_2},$$

be the matrix of (normalized) channel gains between the nodes in S_1 and S_2. Consider first fast fading. Under this assumption, we have

$$C(S_1, S_2) \triangleq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H}) \geq 0: \\ \mathbb{E}(q_{u,u}) \leq 1 \ \forall u \in S_1}} \mathbb{E} \Big(\log \det \big(\boldsymbol{I} + \boldsymbol{H}^\dagger \boldsymbol{Q}(\boldsymbol{H}) \boldsymbol{H} \big) \Big).$$

Define

$$P_{S_1,S_2} \triangleq \sum_{u \in S_1} \sum_{v \in S_2} r_{u,v}^{-\alpha}$$

as the total received power in S_2 from S_1, and set

$$P_{u,S_2} \triangleq P_{\{u\},S_2}$$

with slight abuse of notation. Then

$$\begin{aligned} C(S_1,S_2) &= \max_{\substack{\boldsymbol{Q}(\boldsymbol{H})\geq 0:\\ \mathbb{E}(q_{u,u})\leq P_{u,S_2}\forall u\in S_1}} \mathbb{E}\bigg(\log\det\left(\boldsymbol{I}+\widetilde{\boldsymbol{H}}^\dagger\boldsymbol{Q}(\boldsymbol{H})\widetilde{\boldsymbol{H}}\right)\bigg) \\ &\leq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H})\geq 0:\\ \mathbb{E}(\mathrm{tr}\boldsymbol{Q}(\boldsymbol{H}))\leq P_{S_1,S_2}}} \mathbb{E}\bigg(\log\det\left(\boldsymbol{I}+\widetilde{\boldsymbol{H}}^\dagger\boldsymbol{Q}(\boldsymbol{H})\widetilde{\boldsymbol{H}}\right)\bigg). \end{aligned} \tag{3.32}$$

Define the event

$$B \triangleq \left\{\|\widetilde{\boldsymbol{H}}\|^2 > b\right\}$$

for some b and where $\|\widetilde{\boldsymbol{H}}\|$ denotes the largest singular value of $\widetilde{\boldsymbol{H}}$. In words, B is the event that the channel gains between S_1 and S_2 are "good". We argue that, for appropriately chosen b, the event B has probability zero (i.e., the channel cannot be too "good"). By Markov's inequality

$$\mathbb{P}(B) \leq b^{-m}\mathbb{E}(\|\widetilde{\boldsymbol{H}}\|^{2m}), \tag{3.33}$$

for any m. We continue by upper bounding $\mathbb{E}(\|\widetilde{\boldsymbol{H}}\|^{2m})$. We have

$$\|\widetilde{\boldsymbol{H}}\|^{2k} \leq \mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)$$

for any k, and hence

$$\mathbb{E}(\|\widetilde{\boldsymbol{H}}\|^{2m}) \leq \mathbb{E}\Big(\big(\mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)\big)^{m/k}\Big). \tag{3.34}$$

Now, for any $k \geq m$, we have by Jensen's inequality

$$\mathbb{E}\Big(\big(\mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)\big)^{m/k}\Big) \leq \Big(\mathbb{E}\mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)\Big)^{m/k}. \tag{3.35}$$

Combining (3.33), (3.34), and (3.35) yields

$$\mathbb{P}(B) \leq b^{-m}\Big(\mathbb{E}\mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)\Big)^{m/k} \tag{3.36}$$

for any $k \geq m$.

Now, the arguments in [38, Lemma 5.3] show that

$$\mathbb{E}\big(\mathrm{tr}\big((\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\dagger)^k\big)\big) \leq t_k n\bigg(\max\Big\{1, \max_{v\in S_2}\sum_{u\in S_1}|\tilde{h}_{u,v}|^2\Big\}\bigg)^k,$$

where t_k is the k-th Catalan number. Combining with (3.36), this yields

$$\mathbb{P}(B) \leq \bigg(b^{-1}t_k^{1/k}n^{1/k}\Big(\max\Big\{1, \max_{v\in S_2}\sum_{u\in S_1}|\tilde{h}_{u,v}|^2\Big\}\Big)\bigg)^m.$$

Taking the limit as $k \to \infty$ and using that $t_k^{1/k} \to 4$ yields

$$\mathbb{P}(B) \leq \bigg(b^{-1}4\Big(\max\Big\{1, \max_{v\in S_2}\sum_{u\in S_1}|\tilde{h}_{u,v}|^2\Big\}\Big)\bigg)^m.$$

Assume

$$b > 4\bigg(\max\Big\{1, \max_{v\in S_2}\sum_{u\in S_1}|\tilde{h}_{u,v}|^2\Big\}\bigg), \tag{3.37}$$

then taking the limit as $m \to \infty$ shows that

$$\mathbb{P}(B) = 0.$$

Using this, we can upper bound (3.32) as

$$\begin{aligned}
C(S_1,S_2) &\leq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H})\geq 0:\\ \mathbb{E}(\mathrm{tr}\boldsymbol{Q}(\boldsymbol{H}))\leq P_{S_1,S_2}}} \mathbb{E}\Big(\mathrm{tr}\Big(\widetilde{\boldsymbol{H}}^\dagger\boldsymbol{Q}(\boldsymbol{H})\widetilde{\boldsymbol{H}}\Big)\Big)\\
&= \max_{\substack{\boldsymbol{Q}(\boldsymbol{H})\geq 0:\\ \mathbb{E}(\mathrm{tr}\boldsymbol{Q}(\boldsymbol{H}))\leq P_{S_1,S_2}}} \mathbb{E}\Big(\mathbb{1}_{B^c}\mathrm{tr}\Big(\widetilde{\boldsymbol{H}}^\dagger\boldsymbol{Q}(\boldsymbol{H})\widetilde{\boldsymbol{H}}\Big)\Big)\\
&\leq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H})\geq 0:\\ \mathbb{E}(\mathrm{tr}\boldsymbol{Q}(\boldsymbol{H}))\leq P_{S_1,S_2}}} \mathbb{E}\Big(\mathbb{1}_{B^c}\|\widetilde{\boldsymbol{H}}\|^2\mathrm{tr}\boldsymbol{Q}(\boldsymbol{H})\Big)
\end{aligned}$$

$$\leq bP_{S_1,S_2}.$$

Since this is true for all b satisfying (3.37), we obtain the lemma for the fast fading case.

Under slow fading

$$C(S_1,S_2) \triangleq \max_{\substack{\boldsymbol{Q}\geq 0:\\ q_{u,u}\leq P\ \forall u\in S_1}} \log\det\big(\boldsymbol{I}+\boldsymbol{H}^\dagger\boldsymbol{Q}\boldsymbol{H}\big),$$

and the lemma can be obtained by the same steps. □

We now proceed to the proof of Theorem 3.2. Consider a vertical cut dividing the network into two parts. By the minimum-separation requirement, an area of size $o(n)$ can contain at most $o(n)$ nodes, and hence we can find a cut such that each part is of size $\Theta(n)$ and contains $\Theta(n)$ nodes. Call the left part of the cut S. Since there are $\Theta(n)$ nodes in S and in S^c, there are $\Theta(n)$ sources in S with their destination in S^c with probability $1-o(1)$. For technical reasons we add a node inside each square in $V(n)$ of the form $[id,(i+1)d]\times[jd,(j+1)d]$ for some $i,j\in\mathbb{N}$, where $d\triangleq\sqrt{2\log(n)}$. These additional nodes have no traffic demands on their own, and simply help with the transmission. This can clearly only increase achievable rates. Moreover, this increases the number of nodes in V by less than a factor 2. We now show that

$$C(S,S^c) = O\big(\log^6(n)n^{2-\alpha/2}\big), \tag{3.38}$$

and hence by the cut-set bound [9, Theorem 14.10.1], and since there are $\Theta(n)$ sources in S with their destination in S^c, we have

$$\rho^*(n) = O\big(\log^6(n)n^{1-\alpha/2}\big).$$

We prove (3.38) using Lemma 3.12. To this end, we need to upper bound

$$\max_{v\in S^c}\sum_{u\in S}|\tilde{h}_{u,v}|^2.$$

The proof of [38, Lemma 5.3] shows that if

1. there are less than $\log(n)$ nodes inside $[i, i+1] \times [j, j+1]$ for any $i, j \in \{0, \ldots, \sqrt{n}-1\}$,

2. there is at least one node inside $[id, (i+1)d] \times [jd, (j+1)d]$ for any i, j, where $d \triangleq \sqrt{2 \log n}$,

then

$$\max_{v \in S^c} \sum_{u \in S} |\tilde{h}_{u,v}|^2 \leq K \log^3(n), \tag{3.39}$$

and for $\alpha \in (2, 3]$

$$\sum_{u \in S} \sum_{v \in S^c} r_{u,v}^{-\alpha} \leq \widetilde{K} \log^3(n) n^{2-\alpha/2}, \tag{3.40}$$

for constants $K, \widetilde{K}$. For arbitrary node placement with minimum separation, the first requirement is satisfied for n large enough, since only a constant number of nodes can be contained in each area of constant size. By our addition of nodes into $V(n)$ described above, the second condition is also satisfied. Using Lemma 3.12 with (3.39) and (3.40) yields (3.38), concluding the proof of Theorem 3.2.

3.7 Proof of Theorem 3.3

Consider a node placement with $n/2$ nodes located uniformly on $[0, \sqrt{n}/4] \times [0, \sqrt{n}]$ and $n/2$ nodes located on $[\sqrt{n}/2, \sqrt{n}] \times [0, \sqrt{n}]$ with minimum separation $r_{\min} = 1/2$. A random traffic matrix $\lambda^{\mathsf{UC}}(n)$ is such that at least $n/4$ communication pairs have their sources in the left cluster and destinations in the right cluster with probability $1 - o(1)$. Assume we are dealing with such a $\lambda^{\mathsf{UC}}(n)$ in the following.

In this setup, with multi-hop at least one hop has to cross the gap between the left and the right cluster. Thus, even without any interference from other nodes, we can obtain at most

$$\rho^{\mathsf{MH}}(n) \leq 4^{\alpha} n^{-\alpha/2}.$$

Moreover, considering a cut between the two clusters (say, S and S^c), and applying

Lemma 3.12 yields that

$$\rho^*(n) \leq 16n^{-1}\left(\max\left\{1, \max_{v\in S^c}\sum_{u\in S}|\tilde{h}_{u,v}|^2\right\}\right)\sum_{u\in S}\sum_{v\in S^c} r_{u,v}^{-\alpha}. \tag{3.41}$$

Now note that for any $u \in S$, $v \in S^c$, we have

$$\frac{1}{4}\sqrt{n} \leq r_{u,v} \leq 2\sqrt{n}.$$

Hence

$$\sum_{u\in S}|\tilde{h}_{u,v}|^2 = \sum_{u\in S}\frac{r_{u,v}^{-\alpha}}{\sum_{\tilde{v}\in S^c} r_{u,\tilde{v}}^{-\alpha}} \leq 2^{3\alpha},$$

and

$$\sum_{u\in S}\sum_{v\in S^c} r_{u,v}^{-\alpha} \leq 4^{\alpha-1}n^{2-\alpha/2}.$$

Combining this with (3.41) yields

$$\rho^*(n) \leq 2^{2+5\alpha}n^{1-\alpha/2}$$

for all $\alpha > 2$.

3.8 Proof of Theorem 3.4

We construct a cooperative multi-hop communication scheme and lower bound the per-node rate $\rho^{\mathsf{CMH}}(n)$ it achieves. We use the hierarchical relaying scheme as building block. Assume the node placement $V(n)$ is μ-regular at resolution $d(n)$ for all $n \geq 1$. We show that this implies that we can achieve a per-node rate of at least $d^{3-\alpha}(n)n^{-1/2-\beta(n)}$ as $n \to \infty$. Taking the smallest such $d(n)$ then yields the result.

We consider three cases for the value of $d(n)$ (namely, $d(n) = \Theta(\sqrt{n})$, $d(n) \geq n^{o(1)}$, and $d(n) \leq n^{o(1)}$). First, if $d(n) = \Theta(\sqrt{n})$ as $n \to \infty$ then the result follows directly from Theorem 3.1. Considering a subsequence if necessary, we can therefore assume without loss of generality that $d(n) = o(\sqrt{n})$ in the following.

Second, consider $d(n)$ satisfying

$$d(n) \geq n^{\frac{1}{2+\alpha}\log^{\delta-1/2}(n)}. \tag{3.42}$$

Divide $A(n)$ into squares of sidelength $d(n)$. Since $d(n) = o(\sqrt{n})$, the number of such squares grows unbounded as $n \to \infty$. We now show that we can use multi-hop communication with a hop length of $d(n)$ where each hops is implemented by squares cooperatively sending information to a neighboring square. In other words, we perform cooperative communication at local scale $d(n)$ and multi-hop communication at global scale $\sqrt{n}$.

Since $V(n)$ is μ-regular at resolution $d(n)$, each such square contains at least $\mu d^2(n)$ nodes. Pick the top left most square and construct the square of sidelength $2d(n)$ consisting of it together with its 3 neighbors. Continue in the same fashion, partitioning all of $A(n)$ into squares of sidelength $2d(n)$. Note that each such bigger square contains at least $4\mu d^2(n)$ nodes, and we assume this worst case in the following. Partition $A(n)$ into 4 subsets of those bigger squares such that within each such subset each square is at distance at least $2d(n)$ from any other square (see Figure 3-6). We time share between those 4 subsets. Consider in the following one such subset. For every bigger square, we construct two permutation traffic matrices $\lambda_1^{\mathsf{UC}}(4\mu d^2(n))$ and $\lambda_2^{\mathsf{UC}}(4\mu d^2(n))$. In λ_1^{UC} the nodes in the top two squares have as destinations the nodes in the bottom two squares and the nodes in the bottom two squares have as destinations the nodes in the top two squares (see Figure 3-6). Similarly, λ_2^{UC} contains communication pairs between left and right squares. We time share between λ_1^{UC} and λ_2^{UC}.

Communication according to λ_i^{UC} within bigger squares in the same subset occurs simultaneously. We are going to use hierarchical relaying within each bigger square. This is possible since each such square contains at least $4\mu d^2(n)$ nodes. We have to show that the additional interference from bigger squares in the same subset is such that Theorem 3.1 still applies. In particular, we need to show that the interference has bounded power, say K. Using the same arguments as in the proof of Theorem 3.1

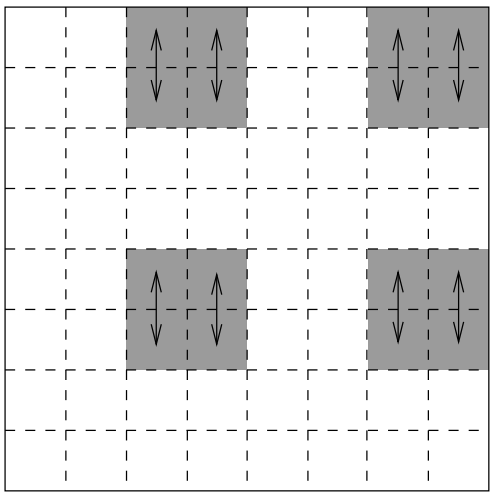

Figure 3-6: Sketch of the construction of the cooperative multi-hop scheme in the proof of Theorem 3.4. The dashed squares have sidelength $d(n)$. The gray area is one of the 4 subsets of bigger squares that communicate simultaneously. The arrows indicate the traffic matrix λ_1^{UC}.

in Section 3.5 yields that this is indeed the case (the interference from other bigger squares here behaves the same way as the interference due to spatial re-use from other active relay subsquares there). With this, we are now dealing with a hierarchical relaying scheme with area $4d^2(n)$, $4\mu d^2(n)$ nodes, and additive noise with power $1+K$. Both the lower number of nodes and the higher noise power will decrease the achievable per-node rate by only some constant factor, and hence Theorem 3.1 shows that under fast fading we can achieve a per-node rate of at least

$$b_1\big(d^2(n)\big)(d^2(n))^{1-\alpha/2} \geq b_1(n)d^{2-\alpha}(n),$$

as $n \to \infty$, where

$$b_1(n) \geq n^{-O\left(\log^{\delta-1/2}(n)\right)}.$$

Moreover, the same rate is achievable under slow fading with probability $1-b_2(d^2(n))$, where

$$b_2(n) \leq \exp\Big(-2^{\Omega\left(\log^{1/2+\delta}(n)\right)}\Big).$$

The setup is the same for all bigger squares within each of the 4 subsets.

We now "shift" the way we defined the bigger squares by $d(n)$ to the right and to the bottom. With this, each new bigger square intersects with 4 bigger squares as defined before. We use the same communication scheme within these new bigger squares and time share between the two ways of defining bigger squares.

Construct now a graph where each vertex corresponds to a square of sidelength $d(n)$ and where two vertices are connected by an edge if they are adjacent in either the same old or new bigger square. This graph is depicted in Figure 3-4 in Section 3.3.

With the above construction, we can communicate along each edge of this graph simultaneously at a per-node rate of

$$\frac{b_1(n)}{16} d^{2-\alpha}(n)$$

in the fast fading case. In the slow fading case, this statement holds with probability at least

$$\begin{aligned} 1 - 4\frac{n}{d^2(n)} b_2(d^2(n)) &\geq 1 - 4\frac{n}{d^2(n)} \exp\Big(- 2^{\Omega\left(\log^{1/2+\delta}(d^2(n))\right)}\Big) \\ &\geq 1 - \exp\Big(K' 2^{\log\log(n)} - 2^{\widetilde{K}\log^{1/2+\delta}(d(n))}\Big) \end{aligned}$$

for constants $K', \widetilde{K}$. By assumption (3.42),

$$\log^{1/2+\delta}\big(d(n)\big) \geq \Big(\frac{1}{2+\alpha}\log^{1/2+\delta}(n)\Big)^{1/2+\delta},$$

and hence

$$1 - \frac{n}{d^2(n)} b_2(d^2(n)) \geq 1 - o(1)$$

as $n \to \infty$, showing that with high probability we achieve the same order rate under slow fading as under fast fading.

The graph constructed forms a grid with $n/d^2(n)$ nodes. Using that each bigger square can contain at most $K_1 d^2(n)$ nodes by the minimum-separation requirement, standard arguments for routing over grid graphs (see [30]) show that in the fast fading

case we can achieve a per-node rate of

$$\rho^{\mathsf{CMH}}(n) \geq \tilde{b}(n) d^{2-\alpha}(n) \frac{d(n)}{\sqrt{n}} \geq \tilde{b}(n) d^{3-\alpha}(n) n^{-1/2},$$

where

$$\tilde{b}(n) = n^{-O\left(\log^{\delta-1/2}(n)\right)}.$$

Moreover, the same statement holds in the slow fading case with probability $1 - o(1)$.

Finally, consider $d(n)$ such that

$$d(n) \leq n^{\frac{1}{2+\alpha}\log^{\delta-1/2}(n)}. \tag{3.43}$$

Construct the same graph as before, but this time we use simple multi-hop communication between adjacent squares of sidelength $d(n)$. By time sharing between the at most $K_1 d^2(n)$ nodes in each square, and since we communicate over a distance of at most $3d(n)$, we achieve under either fast of slow fading a per-node rate between the squares of at least

$$K'' d^{-2-\alpha}(n) \geq K'' n^{-\log^{\delta-1/2}(n)}$$

for some constant K'', and where we have used (3.43). Using the analysis of grid graphs as before, we can achieve a per-node rate of at least

$$\rho^{\mathsf{CMH}}(n) \geq K'' n^{-\log^{\delta-1/2}(n)} \frac{d(n)}{\sqrt{n}} \geq \tilde{b}(n) d^{3-\alpha}(n) n^{-1/2},$$

for either the fast or slow fading case.

3.9 Proof of Theorem 3.5

Consider $V(n)$ with $n/2$ nodes located uniformly on $[0, (\sqrt{n} - d^*(n))/2] \times [0, \sqrt{n}]$ and $n/2$ nodes located uniformly on $[\sqrt{n}/2, \sqrt{n}] \times [0, \sqrt{n}]$ such that $r_{\min} = 1/2$. This node placement is $1/2$-regular at resolution $d^*(n)$. A random traffic matrix $\lambda^{\mathsf{UC}}(n)$ is such that $\Theta(n)$ communication pairs have their sources in the left cluster and destinations

in the right cluster with probability $1 - o(1)$. Assume we are dealing with such a $\lambda^{\mathsf{UC}}(n)$ in the following.

Considering a cut between the two clusters and applying Lemma 3.12 (slightly adapting the arguments in Section 3.6), yields that

$$\rho^*(n) = O\big(\log^6(n) d^{*3-\alpha}(n) n^{-1/2}\big)$$

for $\alpha > 3$.

3.10 Discussion

We briefly discuss several aspects of the proposed hierarchical relaying scheme. Section 3.10.1 comments on the full CSI assumption and Section 3.10.2 on the use of bursty communication. Sections 3.10.3 and 3.10.4 outline how the results obtained here can be extended to the case of dense networks and networks without minimum separation between nodes. Section 3.10.5 compares our hierarchical relaying scheme to the hierarchical cooperation scheme presented in [38]. Section 3.10.6 discusses design guidelines.

3.10.1 Full CSI Assumption

Throughout our analysis, we have made a full CSI assumption. In other words, we assumed that the phase shifts $\{\theta_{u,v}[t]\}_{u,v}$ are available at time t at all nodes in the network. As was pointed out in Section 2.2, a 2-bit quantization of the channel state $\{\theta_{u,v}[t]\}_{u,v}$ available at all nodes at time t is sufficient to obtain the same scaling behavior. This follows by an argument similar to the one used in the analysis of the BC phase in Section 3.4.3, where it is shown that beamforming using a quantized channel state results only in a constant factor rate loss.

3.10.2 Burstiness of Hierarchical Relaying Scheme

The hierarchical relaying scheme presented here is bursty in the sense that nodes communicate at high power during a small fraction of time. This leads to high peak-to-average power ratio, which is undesirable in practice. We chose burstiness in the time domain to simplify the exposition. The same bursty behavior could be achieved in a more practical manner by using CDMA with several orthogonal signatures or by using OFDM with many sub-carriers. Each approach leads to many parallel channels out of which only few are used with higher power. This avoids the issue of burstiness in the time domain.

3.10.3 Dense Networks

Throughout this chapter, we have only considered *extended* networks, i.e, n nodes placed on a square region of area n with a minimum separation of $r_{u,v} \geq r_{\min}$. The results can, however, be recast for *dense* networks, where n nodes are arbitrarily placed on a square region of unit area with a minimum separation of $r_{u,v} \geq r_{\min}/\sqrt{n}$. It suffices to notice that by rescaling power by a factor $n^{-\alpha/2}$ a dense network can essentially be transformed into an extended network with path-loss exponent α (see also [38]). Hence the same result for dense networks can be obtained from the result for extended networks by considering the limit $\alpha \to 2$. Applying this to Theorem 3.1, yields a linear per-node rate scaling of the hierarchical relaying scheme.

3.10.4 Minimum-Separation Requirement

The minimum-separation requirement $r_{\min} \in (0, 1)$ on the node placement is sufficient but not necessary for Theorem 3.1 to hold. A weaker sufficient condition is that a constant fraction of subsquares are dense, as shown in Lemma 3.6 to be a consequence of the minimum-separation requirement. It is straightforward to show that this weaker condition is satisfied with high probability for nodes placed uniformly at random on $[0, \sqrt{n}]^2$. In fact, it can be shown that for a random node placement all subsquares at all levels $0 \leq \ell \leq L(n)$ are dense with high probability. This yields a different proof

of Theorem 5.1 in [38].

3.10.5 Comparison with Prior Work

Both, the hierarchical relaying scheme presented here and the hierarchical scheme presented in [38], share that they use virtual multiple-antenna communication and a hierarchical architecture to achieve essentially global cooperation in the network. The schemes differ, however, in several key aspects, which we point out here.

First, we note that we obtain a slightly better scaling law. Namely

$$b_1(n)n^{1-\alpha/2} \leq \rho^*(n) \leq b_2(n)n^{1-\alpha/2}$$

with

$$\begin{aligned} b_1(n) &\geq n^{-O\left(\log^{\delta-1/2}(n)\right)}, \\ b_2(n) &= O\big(\log^6(n)\big), \end{aligned}$$

for any $\delta \in (0, 1/2)$ obtained here, compared to

$$\tilde{b}_1(n)n^{1-\alpha/2} \leq \rho^*(n) \leq \tilde{b}_2(n)n^{1-\alpha/2}$$

with

$$\begin{aligned} \tilde{b}_1(n) &= \Omega\big(n^{-\varepsilon}\big), \\ \tilde{b}_2(n) &= O\big(n^{\varepsilon}\big), \end{aligned}$$

for any $\varepsilon > 0$ in [38]. For the lower bound (i.e., achievability), this is because the hierarchy here is not of fixed depth L as in [38], but rather of depth $L(n) = \log^{1/2-\delta}(n)$ (for some constant $\delta \in (0, 1/2)$), i.e., changing with n. For the upper bound (i.e., converse), this is due to a sharpening of the arguments in [38].

Second, note that the multi-user decoding at the relay subsquares during the MAC phase and the multi-user encoding during the BC phase are very simple in our

setup. In fact, using matched filter receivers and transmit beamforming, we convert the multi-user encoding and decoding problems into several single-user decoding and encoding problems. This differs from the approach in [38], in which joint decoding of a number of users on the order of the network size is performed. Our results thus imply that these simpler transmitter and receiver structures provide the same scaling as the more complicated joint decoding in [38]. We note that the scheme proposed in [38] can be modified to also use matched filter receivers as suggested here.

Third, and probably most important, the schemes differ in how they achieve the throughput gain from using multiple antennas. In [38], the nodes are located almost regularly with high probability. This allowed the use of a scheme in which a source subsquare directly communicates with a destination subsquare. In other words, the multiple-antenna gain comes from setting up a virtual MIMO channel between the source and the destination. In our setup, the arbitrary location of nodes prevents such an approach. Instead, we use that at least some fixed fraction of subsquares is almost regular (we called them dense subsquares). Source-destination pairs relay their traffic over such a dense subsquare. In other words, the multiple-antenna gain comes from setting up a virtual multiple-antenna MAC and BC. Thus, the hierarchical relaying scheme presented here shows that considerably less structure on the node locations than assumed in [38] suffices to achieve a multiple-antenna gain essentially on the order of the network size. Note also that the additional degree of freedom offered by the choice of relay subsquare for a given source-destination pair makes it possible to extend the result to hold also for slow fading channels.

3.10.6 Design Guidelines

The results presented in this chapter suggest the following design guidelines for communication schemes for large wireless networks. First, in the low path-loss regime, cooperative communication is necessary, and can be achieved regardless of the regularity of the node placement. This cooperative communication is implemented by finding regions in the wireless network that contain many nodes, and in which a hierarchical scheme can be used.

Second, in the high path-loss regime, multi-hop communication should be used whenever the node placement is regular enough for this to be possible. However, for less regular networks, the use of more complicated cooperative communication schemes can be necessary for optimal operation of the network. This is due to large gaps or irregularities in the node placement that make the use of multi-hop communication inefficient.

3.11 Chapter Summary

We considered the problem of the scaling of achievable rates in arbitrary extended wireless networks. We generalized the hierarchical cooperative communication scheme presented in [38] for a fast fading channel model and with random node placements. We proposed a different hierarchical cooperative communication scheme, which also works for arbitrary node placement (with a minimum-separation requirement) and for either fast or slow fading.

For small path-loss exponent $\alpha \in (2, 3]$, we showed that our scheme is order optimal and achieves the same rate irrespective of the node placement. In particular, this rate is equal to the one achievable under random node placement. In other words, the regularity of the node placement has no impact on achievable rates for small path-loss exponent.

The situation is, however, quite different for large path-loss exponent $\alpha > 3$. We argued that in this regime the regularity of the node placement directly impacts the scaling of achievable rates. We then presented a cooperative communication scheme that smoothly "interpolates" between multi-hop and hierarchical cooperative communication depending on the regularity of the node placement. We showed that this scheme is order optimal for all $\alpha > 3$ under adversarial node placement with regularity constraint. This contrasts with the situation for more regular networks (like the ones obtained with high probability through random node placement), in which multi-hop communication is order optimal for all $\alpha > 3$.

Chapter 4

Traffic Heterogeneity

In this chapter, we analyze the scaling of the n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of a wireless network of n randomly placed nodes under a Gaussian fading channel model.

As a first result of this chapter, we present an inner and an outer bound on the unicast capacity region. These bounds coincide in the scaling sense along at least $n^2 - n$ out of n dimensions (corresponding to balanced traffic) in the low path-loss regime $\alpha \in (2, 5]$ and along all n^2 dimensions in the high path-loss regime $\alpha > 5$. These inner and outer bounds approximate the unicast capacity region by a polytope with less than $2n$ faces, each corresponding to a distinct cut (i.e., a subset of nodes) in the wireless network. This polyhedral characterization provides a succinct approximate description of the unicast capacity region even for large values of n. Moreover, it shows that for balanced traffic or $\alpha > 5$ only $2n$ out of 2^n possible cuts in the wireless network are asymptotically relevant and reveals the geometric structure of these relevant cuts.

Second, we establish the approximate equivalence of the wireless network and a wireline tree graph, in the sense that traffic can be transmitted reliably over the wireless channel if and only if approximately the same traffic can be routed over the tree graph. This equivalence is the key component in the derivation of the approximation result for the unicast capacity region and provides insight into the structure of large wireless networks.

Third, we propose a novel three-layer communication architecture that achieves (in the scaling sense) the entire unicast capacity region. The top layer of this scheme treats the wireless network as the aforementioned tree graph and routes messages between sources and their destinations — dealing with heterogeneous traffic demands. The middle layer of this scheme provides this tree abstraction to the top layer by appropriately distributing and concentrating traffic over the wireless network — choosing the level of cooperation in the network. The bottom layer implements this distribution and concentration of messages in the wireless network — dealing with interference and noise. The approximate optimality of this three-layer architecture implies that a separation based approach, in which routing is performed independently of the physical layer, is order-optimal for balanced traffic or in the high path-loss regime $\alpha > 5$.

4.0.1 Organization

The remainder of this chapter is organized as follows. Section 4.1 presents the main results of this chapter. We illustrate the strength of these results in Section 4.2 by analyzing various example scenarios with heterogeneous unicast traffic patterns for which no scaling results were previously known. Section 4.3 provides a high-level description of the proposed communication scheme. Sections 4.4-4.6 contain proofs. Finally, Sections 4.7 and 4.8 contain discussions and concluding remarks.

4.1 Main Results

In this section, we present the main results of this chapter. In Section 4.1.1, we provide inner and outer bounds on the n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network, resulting in a scaling characterization for either most (for $\alpha \in (2, 5]$) or all (for $\alpha > 5$) dimensions of $\Lambda^{\mathsf{UC}}(n)$. In Section 4.1.2, we discuss implications of these results on the behavior of the unicast capacity region for large values of n. In Section 4.1.3, we consider computational aspects.

4.1.1 Unicast Capacity Region

Here we present an inner and outer bound on the n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$. We show that these bounds are tight (in the scaling sense) along n^2-n out of the total n^2 dimensions for $\alpha \in (2,5)$, and are tight (again in the scaling sense) along all n^2 dimensions for $\alpha > 5$.

Our approximate characterization of the unicast capacity region will be given in terms of the total traffic across various regions in the network. To this end, we introduce some notation. Partition $A(n)$ into several square-grids. The ℓ-th square-grid divides $A(n)$ into 4^ℓ squares, each of sidelength $2^{-\ell}\sqrt{n}$, denoted by $\{A_{\ell,i}(n)\}_{i=1}^{4^\ell}$. Let $V_{\ell,i}(n) \subset V(n)$ be the nodes in $A_{\ell,i}(n)$ (see Figure 4-1). The square grids in levels $\ell \in \{1,\ldots,\widetilde{L}(n)\}$ with

$$\widetilde{L}(n) \triangleq \frac{1}{2}\log(n)\big(1-\log^{-1/2}(n)\big), \tag{4.1}$$

will be of particular importance. Note that $\widetilde{L}(n)$ is chosen such that

$$4^{-\widetilde{L}(n)}n = n^{\log^{-1/2}(n)},$$

and hence

$$\lim_{n\to\infty}|A_{\widetilde{L}(n),i}(n)| = \lim_{n\to\infty} 4^{-\widetilde{L}(n)}n = \infty.$$

while at the same time

$$|A_{\widetilde{L}(n),i}(n)| = 4^{-\widetilde{L}(n)}n \leq n^{o(1)},$$

as $n\to\infty$. In other words, the area of the region $A_{\widetilde{L}(n),i}(n)$ at level $\ell = \widetilde{L}(n)$ grows to infinity as $n\to\infty$, but much slower than n.

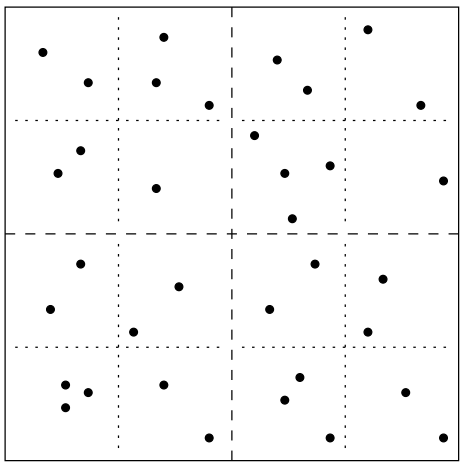

Figure 4-1: Square-grids with $0 \leq \ell \leq 2$. The grid at level $\ell = 0$ is the area $A(n)$ itself. The grid at level $\ell = 1$ is indicated by the dashed lines. The grid at level $\ell = 2$ by the dashed and the dotted lines. Assume for the sake of example that the subsquares are numbered from left to right and then from bottom to top (the precise order of numbering is immaterial). Then $V_{0,1}(n)$ are all the nodes $V(n)$, $V_{1,1}(n)$ are the nine nodes in the lower left corner (separated by dashed lines), and $V_{2,1}(n)$ are the three nodes in the lower left corner (separated by dotted lines).

We are now ready to define the approximate unicast capacity region. Let

$$\begin{aligned}\widehat{\Lambda}_1^{\mathsf{UC}}(n) \triangleq \Big\{\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n\times n} : \sum_{u\in V_{\ell,i}(n)} \sum_{v\notin V_{\ell,i}(n)} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}) \leq (4^{-\ell}n)^{2-\min\{3,\alpha\}/2} \\ \forall \ell \in \{1,\ldots,\widetilde{L}(n)\}, i \in \{1,\ldots,4^\ell\}, \\ \sum_{v\neq u} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}) \leq 1 \\ \forall u \in V(n)\Big\},\end{aligned} \tag{4.2}$$

and

$$\begin{aligned}\widehat{\Lambda}_2^{\mathsf{UC}}(n) \triangleq \Big\{\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n\times n} : \sum_{u\in V_{\ell,i}(n)} \sum_{v\notin V_{\ell,i}(n)} \lambda_{u,v}^{\mathsf{UC}} \leq (4^{-\ell}n)^{2-\min\{3,\alpha\}/2} \\ \forall \ell \in \{1,\ldots,\widetilde{L}(n)\}, i \in \{1,\ldots,4^\ell\}, \\ \sum_{v\neq u} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}) \leq 1 \\ \forall u \in V(n)\Big\}.\end{aligned}$$

$\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ are the collection of all unicast traffic matrices λ^{UC} such that for various cuts $S \subset V(n)$ in the network, the total traffic demand (in either one or both directions)

$$\sum_{u \in S} \sum_{v \notin S} \lambda_{u,v}^{\mathsf{UC}},$$
$$\sum_{u \in S} \sum_{v \notin S} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}),$$

across the cut S is not too big. Note that the number of cuts S we need to consider is actually quite small. In fact, there are at most n sets $\{V_{\ell,i}(n)\}_{\ell,i}$ for $\ell \in \{1, \ldots, \widetilde{L}(n)\}$. Hence $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ are described by $2n$ cuts.

The next theorem shows that $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ is an approximate (in the scaling sense) inner bound and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ is an approximate outer bound to the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network.

Theorem 4.1. *Under either fast or slow fading, for any $\alpha > 2$, there exist*

$$b_1(n) \geq n^{-o(1)},$$
$$b_2(n) = O(\log^6(n)),$$

such that

$$b_1(n)\widehat{\Lambda}_1^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n) \subset b_2(n)\widehat{\Lambda}_2^{\mathsf{UC}}(n),$$

with probability $1 - o(1)$ as $n \to \infty$.

We point out that Theorem 4.1 holds only with probability $1 - o(1)$ for different reasons for the fast and slow fading cases. Under fast fading, the theorem holds only for node placements that are "regular enough". The node placement itself is random, and we show that the required regularity property is satisfied with high probability as $n \to \infty$. Under slow fading, the theorem holds under the same regularity requirements on the node placement, but now it also only holds with high probability for the realization of the fading $\{\theta_{u,v}\}_{u,v}$.

Comparing the expression for $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$, we see that whenever a traffic

matrix λ^{UC} satisfies

$$\sum_{u \in V_{\ell,i}(n)} \sum_{v \notin V_{\ell,i}(n)} (\lambda^{\mathsf{UC}}_{u,v} + \lambda^{\mathsf{UC}}_{v,u}) \leq n^{o(1)} \sum_{u \in V_{\ell,i}(n)} \sum_{v \notin V_{\ell,i}(n)} \lambda^{\mathsf{UC}}_{u,v}, \tag{4.3}$$

for all $\ell \in \{1, \ldots, \widetilde{L}(n)\}$ and $i \in \{1, \ldots 4^\ell\}$, then $\lambda^{\mathsf{UC}} \in \widehat{\Lambda}_2^{\mathsf{UC}}(n)$ implies $n^{-o(1)}\lambda^{\mathsf{UC}} \in \widehat{\Lambda}_1^{\mathsf{UC}}(n)$, and hence for such traffic matrices the inner and outer bounds in Theorem 4.1 coincide in the scaling sense. In particular, this applies for traffic matrices λ^{UC} such that (4.3) holds with equality, and we call such traffic *(approximately) balanced* in the following. Note that the condition of balanced traffic imposes (less than) n linear constraints on λ^{UC}, and hence Theorem 4.1 provides the correct scaling of the n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ along at least $n^2 - n$ dimensions.

In the high path-loss exponent regime ($\alpha > 5$), it can be shown that $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ is also an approximate outer bound.

Theorem 4.2. *Under either fast or slow fading, for any $\alpha > 5$, there exists*

$$b_3(n) = O(\log^6(n)),$$

such that

$$\Lambda^{\mathsf{UC}}(n) \subset b_3(n)\widehat{\Lambda}_1^{\mathsf{UC}}(n),$$

with probability $1 - o(1)$ as $n \to \infty$.

Combining Theorems 4.1 and 4.2 provides a tight scaling characterization for $\alpha > 5$ of the entire unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network as depicted in Figure 4-2. The approximations in both theorems are within a factor $n^{\pm o(1)}$. This factor can be sharpened as is discussed in detail in Section 4.7.2.

4.1.2 Implications of Theorem 4.1 and 4.2

Theorems 4.1 and 4.2 can be applied in two ways. First, the theorems can be used to analyze the asymptotic achievability of a sequence of traffic matrices. Let $\{\lambda^{\mathsf{UC}}(n)\}_{n \geq 1}$

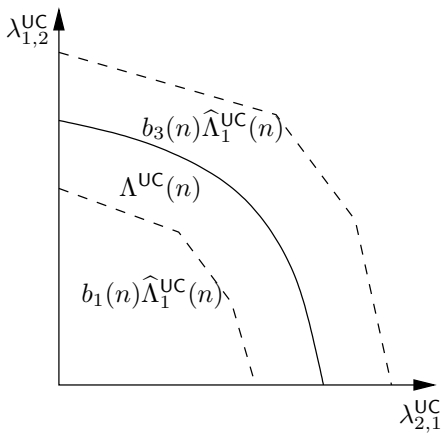

Figure 4-2: For $\alpha > 5$, the set $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ approximates the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network in the sense that $b_1(n)\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ (with $b_1(n) \geq n^{-o(1)}$) provides an inner bound to $\Lambda^{\mathsf{UC}}(n)$ and $b_3(n)\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ (with $b_3(n) = O\big(\log^6(n)\big)$) provides an outer bound to $\Lambda^{\mathsf{UC}}(n)$. The figure shows two dimensions (namely $\lambda_{1,2}^{\mathsf{UC}}$ and $\lambda_{2,1}^{\mathsf{UC}}$) of the n^2-dimensional set $\Lambda^{\mathsf{UC}}(n)$. The same approximation result holds for $\alpha \in (2, 5]$ along at least $n^2 - n$ out of n^2 total dimensions.

be a sequence of unicast traffic matrices with $\lambda^{\mathsf{UC}}(n) \in \mathbb{R}_+^{n \times n}$. Define

$$\rho^*_{\lambda^{\mathsf{UC}}}(n) \triangleq \sup\{\rho : \rho\lambda^{\mathsf{UC}}(n) \in \Lambda^{\mathsf{UC}}(n)\},$$
$$\hat{\rho}^*_{\lambda^{\mathsf{UC}}}(n) \triangleq \sup\{\hat{\rho} : \hat{\rho}\lambda^{\mathsf{UC}}(n) \in \widehat{\Lambda}_1^{\mathsf{UC}}(n)\},$$

i.e., $\rho^*_{\lambda^{\mathsf{UC}}}(n)$ is the largest multiplier ρ such that the scaled traffic matrix $\rho\lambda^{\mathsf{UC}}(n)$ is contained in $\Lambda^{\mathsf{UC}}(n)$ (and similar for $\hat{\rho}^*_{\lambda^{\mathsf{UC}}}(n)$ with respect to $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$). Then Theorems 4.1 and 4.2 provide asymptotic information about the achievability of $\{\lambda^{\mathsf{UC}}(n)\}_{n\geq 1}$ in the sense that if either the $\lambda^{\mathsf{UC}}(n)$ are balanced or if $\alpha > 5$ then[1]

$$\lim_{n\to\infty} \frac{\log(\rho^*_{\lambda^{\mathsf{UC}}}(n))}{\log(n)} = \lim_{n\to\infty} \frac{\log(\hat{\rho}^*_{\lambda^{\mathsf{UC}}}(n))}{\log(n)}.$$

Second, Theorems 4.1 and 4.2 provide information about the shape of the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$. We now argue that even though the approximation $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ of $\Lambda^{\mathsf{UC}}(n)$ is only up to $n^{o(1)}$ scaling, its shape is largely preserved.

[1]We assume here that the limits exist, otherwise the same statement holds for lim sup and lim inf.

To illustrate this point, consider a rectangle

$$R(n) \triangleq [0, r_1(n)] \times [0, r_2(n)],$$

and let

$$\widehat{R}(n) \triangleq [0, \hat{r}_1(n)] \times [0, \hat{r}_2(n)],$$

where

$$\hat{r}_i \triangleq b_i(n) r_i(n)$$

for some $b_i(n) = n^{\pm o(1)}$, be its approximation. The shape of $R(n)$ is then determined by the ratio between $r_1(n)$ and $r_2(n)$. For example, assume $r_1(n) = n^{\beta} r_2(n)$. Then

$$\frac{\hat{r}_1(n)}{\hat{r}_2(n)} = n^{\beta \pm o(1)} = n^{\pm o(1)} \frac{r_1(n)}{r_2(n)},$$

i.e.,

$$\lim_{n\to\infty} \frac{\log\big(r_1(n)/r_2(n)\big)}{\log(n)} = \beta = \lim_{n\to\infty} \frac{\log\big(\hat{r}_1(n)/\hat{r}_2(n)\big)}{\log(n)},$$

and hence the approximation $\widehat{R}(n)$ preserves the exponent of the ratio of sidelengths of $R(n)$. In other words, if the two sidelengths $r_1(n)$ and $r_2(n)$ differ on exponential scale (i.e., by a factor n^{β} for $\beta \neq 0$) then this shape information is preserved by the approximation $\widehat{R}(n)$.

Let us now return to the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ and its approximation $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$. We consider several boundary points of $\Lambda^{\mathsf{UC}}(n)$ and show that their behavior varies at scale n^{β} for different values of β. From the discussion in the previous paragraph, this implies that a significant part of the shape of $\Lambda^{\mathsf{UC}}(n)$ is preserved by its approximation $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$. First, let $\lambda^{\mathsf{UC}} \triangleq \rho(n)\mathbf{1}$ for some scalar $\rho(n)$ depending only on n, and where $\mathbf{1}$ is the $n \times n$ matrix of all ones. If $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$ then the largest achievable value of $\rho(n)$ is $\rho^*(n) \leq n^{-\min\{3,\alpha\}/2+o(1)}$ (by applying Theorem 4.1). Second, let λ^{UC} such that $\lambda^{\mathsf{UC}}_{u^*,w^*} = \rho(n)$ for only one source-destination pair (u^*, w^*) with $u^* \neq w^*$ and $\lambda^{\mathsf{UC}}_{u,w} = 0$ otherwise. Then $\rho^*(n)$, the largest achievable value of $\rho(n)$, satisfies $\rho^*(n) \geq n^{-o(1)}$ (again by Theorem 4.1). Hence the boundary

points of $\Lambda^{\mathsf{UC}}(n)$ vary at least from $n^{-\min\{3,\alpha\}/2+o(1)}$ to $n^{-o(1)}$, and this variation on exponential scale is preserved by $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$.

4.1.3 Computational Aspects

Since we are interested in large wireless networks, computational aspects are of importance. In this section, we show that the region $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ can be efficiently described. By Theorems 4.1 and 4.2, this provides a computationally efficient approximate description of the entire unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ for $\alpha > 5$ and of $n^2 - n$ of its n^2 dimensions for $\alpha \in (2, 5]$.

Note that $\Lambda^{\mathsf{UC}}(n)$ is a n^2-dimensional set, and hence its shape could be rather complex. In particular, in the special cases where the capacity region is known, its description is often in terms of cut-set bounds. Since there are 2^n possible subsets of n nodes, there are 2^n possible cut-set bounds to be considered. In other words, the description complexity of $\Lambda^{\mathsf{UC}}(n)$ is likely to be growing exponentially in n. On the other hand, as was pointed out in Section 4.1.1, the description of $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ is in terms of only $2n$ cuts. This implies that $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ can be computed efficiently (i.e., in polynomial time in n). Hence even though the description complexity of $\Lambda^{\mathsf{UC}}(n)$ is likely to be of order $\Theta(2^n)$, the description complexity of its approximation $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ is only of order $\Theta(n)$ — an exponential reduction. In particular, consider a unicast traffic matrix λ^{UC} and assume that either $\alpha > 5$ or that λ^{UC} is balanced, then this implies that membership $\lambda^{\mathsf{UC}} \in \widehat{\Lambda}^{\mathsf{UC}}(n)$ (and hence by Theorems 4.1 and 4.2 also the approximate achievability of the unicast traffic matrix λ^{UC}) can be computed in polynomial time in the network size n. More precisely, evaluating each of the $\Theta(n)$ cuts takes at most $\Theta(n^2)$ operations, yielding a $\Theta(n^3)$-time algorithm for approximate testing of membership in $\Lambda^{\mathsf{UC}}(n)$.

4.2 Example Scenarios

We next illustrate the strength of the above results by determining achievable rates in a few specific wireless network scenarios with non-uniform traffic patterns. We

illustrate the impact on achievable rates of various sources of traffic heterogeneities — variation of distance between source-destination pairs, variation of amount of traffic between different pairs, sources with multiple destinations.

Example 4.1. *Multiple classes of source-destination pairs*

There are K classes of source-destination pairs, for some fixed K. Each source node in class i generates traffic at the same rate $\rho_i(n)$ for a destination node that is chosen randomly within distance $\Theta(n^{\beta_i/2})$, for some fixed $\beta_i \in [0,1]$. Each node randomly picks the class it belongs to. The resulting traffic matrix is approximately balanced with high probability, and applying Theorem 4.1 shows that $\rho_i^*(n)$, the largest achievable value of $\rho_i(n)$, satisfies

$$\rho_i^*(n) = n^{\beta_i(1-\bar{\alpha}/2)\pm o(1)},$$

with probability $1 - o(1)$ for all i, and where

$$\bar{\alpha} \triangleq \min\{3, \alpha\}.$$

Hence, for a fixed number of classes K, source nodes in each class can obtain rates as a function of only the source-destination separation in that class.

Set $\tilde{n}_i \triangleq n^{\beta_i}$, and note that $\tilde{n}_i$ is on the order of the expected number of nodes that are closer to a source than its destination. Then

$$\rho_i^*(n) = n^{\pm o(1)} \tilde{n}_i^{1-\bar{\alpha}/2}.$$

Now $\tilde{n}_i^{1-\bar{\alpha}/2}$ is precisely the per-node rate that is achievable for an extended network with $\tilde{n}_i$ nodes under random source-destination pairing [38]. In other words, the local traffic pattern here allows us to obtain a rate that is as good as the one achievable under random source-destination pairing for a much smaller network. $\Diamond$

Example 4.2. *Traffic variation with source-destination separation*

Assume each node is source for exactly one destination chosen uniformly at random from among all the other nodes (as in the traditional setting). However, instead of all

sources generating traffic at the same rate, source node u generates traffic at a rate that is a function of its separation from destination w, i.e., the traffic matrix is given by $\lambda_{u,w}^{\mathsf{UC}} = \psi(r_{u,w})$ for some function ψ. In particular, let us consider

$$\psi(r) \triangleq \rho(n) \times \begin{cases} r^{\beta} & \text{if } r \geq 1, \\ 1 & \text{else} \end{cases}$$

for some fixed $\beta \in \mathbb{R}$ and some $\rho(n)$ depending only on n. The traditional setting corresponds to $\beta = 0$, in which case all n source-destination pairs communicate at uniform rate. Such traffic is approximately balanced with high probability, and using Theorem 4.1 establishes the scaling of $\rho^*(n)$, the largest achievable value of $\rho(n)$, as

$$\rho^*(n) = \begin{cases} n^{1-(\bar{\alpha}+\beta)/2 \pm o(1)} & \text{if } \beta \geq 2 - \bar{\alpha}, \\ n^{\pm o(1)} & \text{else,} \end{cases}$$

with probability $1 - o(1)$. For $\beta = 0$, and noting that $2 - \bar{\alpha} \leq 0$, this recovers the results from [38] for random source-destination pairing with uniform rate. $\Diamond$

Example 4.3. *Source-destination separation variation*

Each node generates traffic at the same rate $\rho(n)$. For each source u we pick one destination $w(u)$ independently at random at distance s with density $\psi(s)$, i.e., for all $r \in [0, \sqrt{n}]$

$$\mathbb{P}(r_{u,w(u)} \leq r | V(n)) \approx \frac{1}{Z(n)} \int_{s=1}^{r} \psi(s) ds, \tag{4.4}$$

for

$$\psi(r) \triangleq r^{\beta},$$

for some fixed $\beta \in \mathbb{R}$, and for normalization constant

$$Z(n) \triangleq \int_{s=1}^{\sqrt{n}} \psi(s) ds$$

(the relation is only approximate since the number of nodes is finite). Note that the node placement $V(n)$ in (4.4) is fixed, and hence so are $r_{u,v}$ for all pairs $u, v \in V(n)$.

The randomness in (4.4) is due to the random choice of destination $w(u)$ for source u. Note also that the traditional setup of choosing destinations uniformly at random from among all other nodes corresponds essentially to $\beta = 1$. Finally, note that this traffic is approximately balanced with high probability. The scaling of $\rho^*(n)$, the largest achievable $\rho(n)$, is thus given by Theorem 4.1 as

$$\rho^*(n) = \begin{cases} n^{1-\bar{\alpha}/2 \pm o(1)} & \text{if } \beta \geq -1, \\ n^{(1-\bar{\alpha}-\beta)/2 \pm o(1)} & \text{if } 1-\bar{\alpha} \leq \beta < -1, \\ n^{\pm o(1)} & \text{else,} \end{cases}$$

with probability $1 - o(1)$. For $\beta = 1$ this coincides again with the results from [38] for random source-destination pairing with uniform rate. ◊

Example 4.4. *Sources with multiple destinations*

All the example scenarios so far are concerned with traffic in which each node is source exactly once. Here we consider more general traffic patterns. There are K classes of source nodes, for some fixed K. Each source node in class i has $\Theta(n^{\beta_i})$ destination nodes for some fixed $\beta_i \in [0, 1]$ and generates independent traffic at the same rate $\rho_i(n)$ for each of them (i.e., we still consider unicast traffic). Each of these destination nodes is chosen uniformly at random among the $n-1$ other nodes. Every node randomly picks the class it belongs to. Noting that the resulting traffic matrix is approximately balanced with high probability, Theorem 4.1 provides the following scaling of the rates achievable by different classes:

$$\rho_i^*(n) = n^{1-\beta_i-\bar{\alpha}/2 \pm o(1)},$$

with probability $1 - o(1)$ for all i. In other words, for each source node time sharing between all K classes and then (within each class) between all its $\Theta(n^{\beta_i})$ destination nodes is order-optimal in this scenario. However, different sources are operating simultaneously. ◊

4.3 Communication Scheme for Unicast Traffic

In this section, we provide a high-level description of the communication scheme used to prove achievability (i.e., the inner bound) in Theorem 4.1. This scheme has a tree structure, that makes it convenient to work with. This tree structure is crucial in proving the compact approximation of the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ in Theorems 4.1 and 4.2. The communication scheme for general unicast traffic uses as a building block the multi-hop and hierarchical relaying schemes for random source-destination pairing with uniform rate discussed in Chapter 3.

The communication scheme consists of three layers: A top or routing layer, a middle or cooperation layer, and a bottom or physical layer. The routing layer of this scheme treats the wireless network as a tree graph G and routes messages between sources and their destinations — dealing with heterogeneous traffic demands. The cooperation layer of this scheme provides this tree abstraction G to the top layer by appropriately distributing and concentrating traffic over the wireless network — choosing the level of cooperation in the network. The physical layer implements this distribution and concentration of messages in the wireless network — dealing with interference and noise.

Seen from the routing layer, the network consists of a noiseless capacitated[2] graph G. This graph is a tree, whose leaf nodes are the nodes $V(n)$ in the wireless network. The internal nodes of G represent larger clusters of nodes (i.e., subsets of $V(n)$) in the wireless network. More precisely, each internal node in G represents a set $V_{\ell,i}(n)$ for $\ell \in \{1, \ldots, \widetilde{L}(n)\}$ and $i \in \{1, \ldots, 4^\ell\}$. Consider two sets $V_{\ell,i}(n), V_{\ell+1,j}(n)$ and let ν, μ be the corresponding internal nodes in G. Then ν and μ are connected by an edge in G if $V_{\ell+1,j}(n) \subset V_{\ell,i}(n)$. Similarly, for $V_{\widetilde{L}(n),i}(n)$ and corresponding internal node ν in G, a leaf node u in G is connected by an edge to ν if $u \in V_{\widetilde{n},i}(n)$ (recall that the leaf nodes of G are the nodes $V(n)$ in the wireless network). This construction is shown in Figure 4-3. In the routing layer, messages are sent from each source to its destination by routing them over G. To send information along an edge of G, the

[2]A graph $G = (V_G, E_G)$ is *capacitated* if every edge $e \in E_G$ is associated with a nonnegative capacity c_e.

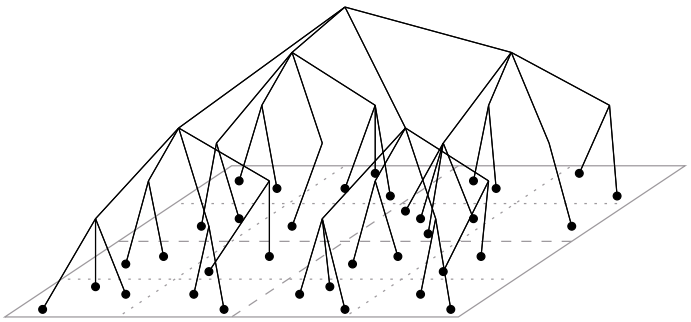

Figure 4-3: Construction of the tree graph G. We consider the same nodes as in Figure 4-1 with $\widetilde{L}(n) = 2$. The leaves of G are the nodes $V(n)$ of the wireless network. They are always at level $\ell = \widetilde{L}(n) + 1$ (i.e., 3 in this example). At level $0 \leq \ell \leq \widetilde{L}(n)$ in G, there are 4^ℓ nodes. The tree structure is the one induced by the grid decomposition $\{V_{\ell,i}(n)\}_{\ell,i}$ as shown in Figure 4-1. Level 0 contains the root node of G.

routing layer calls upon the cooperation layer.

The cooperation layer implements the tree abstraction G. This is done by ensuring that whenever a message is located at a node in G, it es evenly distributed over the corresponding cluster in the wireless network, i.e., every node in the cluster has access to a distinct part of equal length of the message. To send information from a child node to its parent in G (i.e., towards the root node of G), the message at the cluster in $V(n)$ represented by the child node is distributed evenly among all nodes in the bigger cluster in $V(n)$ represented by the parent node. More precisely, let ν be a child node of μ in G, and let $V_{\ell+1,i}(n), V_{\ell,j}(n)$ the corresponding subsets of $V(n)$. Consider the cooperation layer being called by the routing layer to send a message from ν to its parent μ over G. In the wireless network, we assume each node in $V_{\ell+1,i}(n)$ has access to a distinct $1/|V_{\ell+1,i}(n)|$ fraction of the message to be sent. Each node in $V_{\ell+1,i}(n)$ splits its message part into four distinct parts of equal length. It keeps one part for itself and sends the other three parts to three nodes in $V_{\ell,j}(n) \setminus V_{\ell+1,i}(n)$. After each node in $V_{\ell+1,i}(n)$ has sent their message parts, each node in $V_{\ell,j}(n)$ now as access to a distinct $1/|V_{\ell,j}(n)|$ fraction of the message. To send information from a parent node to a child node in G (i.e., away from the root node of G), the message at the cluster in $V(n)$ represented by the parent node is concentrated on the cluster in $V(n)$ represented by the child node. More precisely, consider the same nodes ν

and μ in G corresponding to $V_{\ell+1,i}(n)$ and $V_{\ell,j}(n)$ in $V(n)$. Consider the cooperation layer being called by the routing layer to send a message from μ to its child ν. In the wireless network, we assume each node in $V_{\ell,j}(n)$ has access to a distinct $1/|V_{\ell,j}(n)|$ fraction of the message to be sent. Each node in $V_{\ell,j}(n)$ sends its message part to another node in $V_{\ell+1,i}(n)$. After each node in $V_{\ell,j}(n)$ has sent their message part, each node in $V_{\ell+1,i}(n)$ now as access to a distinct $1/|V_{\ell+1,i}(n)|$ fraction of the message. To implement this distribution and concentration of messages, the cooperation layer calls upon the physical layer.

The physical layer performs the distribution and concentration of messages. Note that the traffic induced by the cooperation layer in the physical layer is very uniform, and closely resembles a permutation traffic (in which each node in the wireless network is source and destination once and all these source-destination pairs want to communicate at equal rate). Hence we can use either cooperative communication (for $\alpha \in (2, 3]$) or multi-hop communication (for $\alpha > 3$) for the transmission of this traffic. See Chapter 3 for a detailed description of these two schemes. It is this operation in the physical layer that determines the edge capacities of the graph G as seen from the routing layer.

The operation of this three-layer architecture is illustrated in the following example.

Example 4.5. Consider a single source-destination pair (u, v). The corresponding operation of the three-layer architecture is depicted in Figure 4-4.

In the routing layer, the message is routed over the tree graph G between u and w (indicated in black in the figure). The middle plane in the figure shows the induced behavior from using the second edge along this path (indicated in solid black in the figure) in the cooperation layer. The bottom plane in the figure shows (part of) the corresponding actions induced in the physical layer. Let us now consider the specific operations of the three layers for the single message between u and v. Since G is a tree, there is a unique path between u and v, and the routing layer sends the message over the edges along this path. Consider now the first such edge. Using this edge in the routing layer induces the following actions in the cooperation layer. The node u,

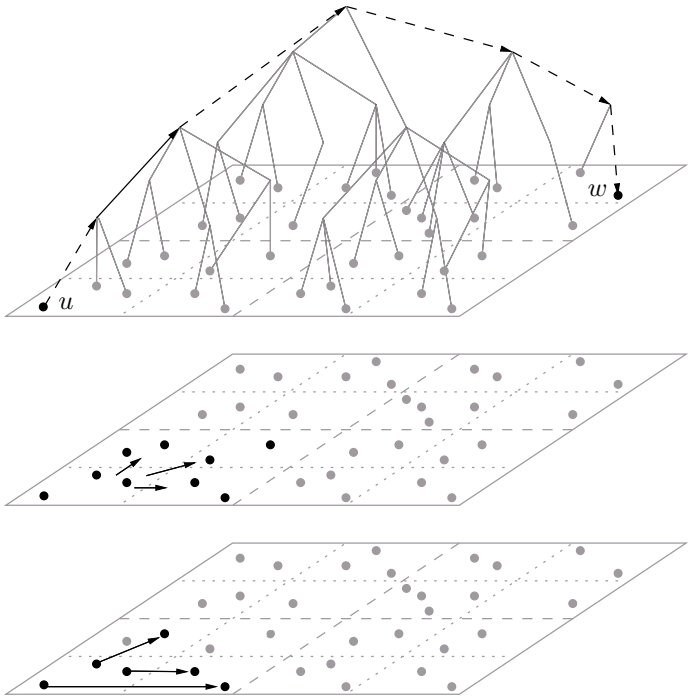

Figure 4-4: Example operation of the three-layer architecture under unicast traffic. The three layers depicted are (from top to bottom in the figure) the routing layer, the cooperation layer, and the physical layer.

having access to the entire message, splits that message into 3 distinct parts of equal length. It keeps one part, and sends the other two parts to the two other nodes in $V_{2,1}(n)$ (i.e., lower left square at level $\ell = 2$ in the hierarchy). In other words, after the message has traversed the edge between u and its parent node in the routing layer, all nodes in $V_{2,1}(n)$ in the cooperation layer have access to a distinct $1/3$ fraction of the original message. The edges in the routing layer leading up the tree (i.e., towards the root node) are implemented in the cooperation layer in a similar fashion by further distributing the message over the wireless network. By the time the message reaches the root node of G in the routing layer, the cooperation layer has distributed the message over the entire network and every node in $V(n)$ has access to a distinct $1/n$ fraction of the original message. Communication down the tree in the routing layer is implemented in the cooperation layer by concentrating messages over smaller regions in the wireless network. To physically perform this distribution and concentration of messages, the cooperation layer calls upon the physical layer, which uses either hierarchical relaying or multi-hop communication. $\Diamond$

4.4 Auxiliary Lemmas

In this section, we provide auxiliary results, which will be used several times in the following. These results are grouped into three parts. In Section 4.4.1, we describe regularity properties exhibited with high probability by the random node placement. In Section 4.4.2, we provide auxiliary upper bounds on the performance of any scheme in terms of cut-set bounds. Finally, in Section 4.4.3, we describe auxiliary results on the performance of hierarchical relaying and multi-hop communication.

4.4.1 Regularity Lemmas

Here we prove several regularity properties that are satisfied with high probability by a random node placement. Formally, define $\mathcal{V}(n)$ to be the collection of all node

placements $V(n)$ that satisfy the following conditions:

$$
\begin{aligned}
r_{u,v} &> n^{-1} && \forall u, v \in V(n), \\
\big|V_{\ell,i}(n)\big| &\leq \log(n) && \text{for } \ell = \frac{1}{2}\log(n) \text{ and } \forall i \in \{1, \ldots, 4^{\ell}\}, \\
\big|V_{\ell,i}(n)\big| &\geq 1 && \text{for } \ell = \frac{1}{2}\log\Big(\frac{n}{2\log(n)}\Big) \text{ and } \forall i \in \{1, \ldots, 4^{\ell}\}, \\
\big|V_{\ell,i}(n)\big| &\in [4^{-\ell-1}n, 4^{-\ell+1}n] \quad \forall \ell \in \Big\{1, \ldots, \frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big)\Big\}, i \in \{1, \ldots, 4^{\ell}\}.
\end{aligned}
$$

The first condition is that the minimum separation between node pairs is not too small. The second condition is that all squares of area 1 contain at most $\log(n)$ nodes. The third condition is that all squares of area $2\log(n)$ contain at least one node. The fourth condition is that all squares up to level $\frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big)$ contain a number of nodes proportional to their area. Note that, since

$$\widetilde{L}(n) \leq \frac{1}{2}\log(n)\big(1 - \log^{5/6}(n)\big),$$

this holds in particular for nodes up to level $\widetilde{L}(n)$. The goal of this section is to prove that

$$\mathbb{P}(V(n) \in \mathcal{V}(n)) = 1 - o(1),$$

as $n \to \infty$.

The first lemma shows that the minimum separation in a random node placement is at least n^{-1} with high probability.

Lemma 4.3.

$$\mathbb{P}\Big(\min_{u \in V(n), v \in V(n) \setminus \{u\}} r_{u,v} > n^{-1}\Big) = 1 - o(1),$$

as $n \to \infty$.

Proof. For $u, v \in V$, let

$$B_{u,v} \triangleq \{r_{u,v} \leq r\}$$

for some r (depending only on n). Fix a node $u \in V$, then

$$\mathbb{P}(B_{u,v}|u) \leq \frac{r^2\pi}{n},$$

(the inequality being due to boundary effects). Moreover, the events $\{B_{u,v}\}_{v\in V\setminus\{u\}}$ are independent conditioned on u, and thus

$$\mathbb{P}\Big(\cap_{v\in V\setminus\{u\}} B^c_{u,v}\big|u\Big) = \prod_{v\in V\setminus\{u\}} \mathbb{P}(B^c_{u,v}|u) \geq \Big(1-\frac{r^2\pi}{n}\Big)^n.$$

From this,

$$\begin{aligned}
\mathbb{P}\Big(\min_{u\in V, v\in V\setminus\{u\}} r_{u,v} \leq r\Big) &= \mathbb{P}\Big(\cup_{u\in V, v\in V\setminus\{u\}} B_{u,v}\Big) \\
&\leq \sum_{u\in V} \mathbb{P}\Big(\cup_{v\in V\setminus\{u\}} B_{u,v}\Big) \\
&= \sum_{u\in V}\Big(1-\mathbb{P}\Big(\cap_{v\in V\setminus\{u\}} B^c_{u,v}\Big)\Big) \\
&= \sum_{u\in V}\Big(1-\mathbb{E}\Big(\mathbb{P}\Big(\cap_{v\in V\setminus\{u\}} B^c_{u,v}\big|u\Big)\Big)\Big) \\
&\leq \sum_{u\in V}\Big(1-\Big(1-\frac{r^2\pi}{n}\Big)^n\Big) \\
&= n\Big(1-\Big(1-\frac{r^2\pi}{n}\Big)^n\Big),
\end{aligned}$$

which converges to zero for $r = n^{-1}$. □

The next lemma asserts that if $\tilde{\ell}(n)$ is not too large then all squares $\{V_{\ell,i}(n)\}_{\ell,i}$ for $\ell \in \{1,\ldots,\tilde{\ell}(n)\}$ and $i \in \{1,\ldots,4^\ell\}$ in the grid decomposition of $V(n)$ contain a number of nodes that is proportional to their area.

Lemma 4.4. *If $\tilde{\ell}(n)$ satisfies*

$$\lim_{n\to\infty} \frac{\tilde{\ell}(n)}{4^{-\tilde{\ell}(n)}n} = 0$$

then

$$\mathbb{P}\Big(\bigcap_{\ell=1}^{\tilde{\ell}(n)}\bigcap_{i=1}^{4^\ell}\big\{|V_{\ell,i}(n)|\in[4^{-\ell-1}n,4^{-\ell+1}n]\big\}\Big)=1-o(1)$$

as $n\to\infty$. In particular, this holds for

$$\tilde{\ell}(n)=\frac{1}{2}\log(n)\big(1-\log^{-5/6}(n)\big),$$

and for $\tilde{\ell}(n)=\widetilde{L}(n)$.

Proof. Let B_u be the event that node u lies in $A_{\ell,i}$ for fixed ℓ,i. Note that

$$\sum_{u\in V}\mathbb{1}_{B_u}=|V_{\ell,i}|$$

by definition, and that

$$\mathbb{P}(B_u)=4^{-\ell}.$$

Hence using the Chernoff bound

$$\mathbb{P}\Big(\sum_{u\in V}\mathbb{1}_{B_u}\notin[4^{-\ell-1}n,4^{-\ell+1}n]\Big)\le\exp(-K4^{-\ell}n),$$

for some constant K. From this, we obtain for $\ell=\tilde{\ell}(n)$,

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i=1}^{4^{\tilde{\ell}(n)}}\big\{|V_{\tilde{\ell}(n),i}|\in[4^{-\tilde{\ell}(n)-1}n,4^{-\tilde{\ell}(n)+1}n]\big\}\Big)&\\ \ge 1-\sum_{i=1}^{4^{\tilde{\ell}(n)}}\mathbb{P}\big(|V_{\tilde{\ell}(n),i}|\notin[4^{-\tilde{\ell}(n)-1}n,4^{-\tilde{\ell}(n)+1}n]\big)&\qquad(4.5)\\ \ge 1-4^{\tilde{\ell}(n)}\exp(-K4^{-\tilde{\ell}(n)}n)&\\ \ge 1-\exp(\widetilde{K}\tilde{\ell}(n)-K4^{-\tilde{\ell}(n)}n),&\end{aligned}$$

for some constant $\widetilde{K}$. By assumption

$$\lim_{n\to\infty}\frac{\tilde{\ell}(n)}{4^{-\tilde{\ell}(n)}n}=0,$$

and hence

$$\mathbb{P}\Big(\bigcap_{i=1}^{4^{\tilde{\ell}(n)}} \big\{|V_{\tilde{\ell}(n),i}| \in [4^{-\tilde{\ell}(n)-1}n, 4^{-\tilde{\ell}(n)+1}n]\big\}\Big) \geq 1 - o(1),$$

as $n \to \infty$. Since the $\{A_{\ell,i}\}_{\ell,i}$ are nested as a function of ℓ, we have

$$\bigcap_{\ell=1}^{\tilde{\ell}(n)} \bigcap_{i=1}^{4^{\ell}} \big\{|V_{\ell,i}| \in [4^{-\ell-1}n, 4^{-\ell+1}n]\big\} = \bigcap_{i=1}^{4^{\tilde{\ell}(n)}} \big\{|V_{\tilde{\ell}(n),i}| \in [4^{-\tilde{\ell}(n)-1}n, 4^{-\tilde{\ell}(n)+1}n]\big\},$$

which, combined with (4.5), proves the first part of the lemma.

For the second part, note that for

$$\tilde{\ell}(n) = \frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big),$$

we have

$$\begin{aligned}\frac{\tilde{\ell}(n)}{4^{-\tilde{\ell}(n)}n} &= \frac{\frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big)}{2^{\log^{1/6}(n)}} \\ &\leq \frac{\log(n)}{2^{\log^{1/6}(n)}} \\ &= 2^{\log\log(n) - \log^{1/6}(n)} \to 0,\end{aligned}$$

and hence the lemma is valid in this case. The same holds for $\tilde{\ell}(n) = \widetilde{L}(n)$ since

$$\widetilde{L}(n) \leq \frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big).$$

□

We are now ready to prove that a random node placement $V(n)$ is in $\mathcal{V}(n)$ with high probability as $n \to \infty$ (i.e., is fairly "regular" with high probability).

Lemma 4.5.

$$\mathbb{P}(V(n) \in \mathcal{V}(n)) = 1 - o(1),$$

as $n \to \infty$.

Proof. The first condition,

$$r_{u,v} > n^{-1} \qquad \forall u, v \in V,$$

holds with probability $1 - o(1)$ by Lemma 4.3. The second and third conditions,

$$\begin{aligned} \left|V_{\ell,i}\right| &\leq \log(n) && \text{for } \ell = \frac{1}{2}\log(n) \text{ and } \forall i \in \{1, \ldots, 4^\ell\}, \\ \left|V_{\ell,i}\right| &\geq 1 && \text{for } \ell = \frac{1}{2}\log\left(\frac{n}{2\log(n)}\right) \text{ and } \forall i \in \{1, \ldots, 4^\ell\}, \end{aligned}$$

are shown in [38, Lemma 5.1] to hold with probability $1 - o(1)$. The fourth condition,

$$\left|V_{\ell,i}\right| \in [4^{-\ell-1}n, 4^{-\ell+1}n] \quad \forall \ell \in \left\{1, \ldots, \frac{1}{2}\log(n)\left(1 - \log^{-5/6}(n)\right)\right\}, i \in \{1, \ldots, 4^\ell\},$$

holds with probability $1 - o(1)$ by Lemma 4.4. Together, this proves the desired result. □

4.4.2 Converse Lemmas

Here we prove several auxiliary converse results. The first lemma bounds the maximal achievable sum rate for every individual node (i.e., the total traffic for which a fixed node is either source or destination).

Lemma 4.6. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b(n) = O(\log(n))$ such that for all $V(n) \in \mathcal{V}(n)$, $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$, $u \in V(n)$,*

$$\sum_{v \in V(n) \setminus \{u\}} \lambda^{\mathsf{UC}}_{u,v} \leq b(n), \tag{4.6}$$

$$\sum_{v \in V(n) \setminus \{u\}} \lambda^{\mathsf{UC}}_{v,u} \leq b(n). \tag{4.7}$$

Proof. The argument follows the one in [38, Theorem 3.1]. For any $S_1, S_2 \subset V$, denote by $C(S_1, S_2)$ the MIMO capacity between nodes in S_1 and nodes in S_2. Consider first

(4.6). By the cut-set bound [9, Theorem 14.10.1],

$$\sum_{v \neq u} \lambda_{u,v}^{\mathsf{UC}} \leq C(\{u\}, \{u\}^c).$$

Here $C(\{u\}, \{u\}^c)$ is the SIMO capacity between u and the nodes in $\{u\}^c$, i.e.,

$$\begin{aligned} C(\{u\}, \{u\}^c) &= \log\left(1 + \textstyle\sum_{v \neq u} |h_{u,v}|^2\right) \\ &\leq \log(1 + (n-1)n^{\alpha}) \\ &\leq K \log(n), \end{aligned}$$

with

$$K \triangleq 2 + \alpha,$$

and where for the first inequality we have used that since $V \in \mathcal{V}$, we have $r_{u,v} \geq n^{-1}$ for all $u, v \in V$.

Similarly, for (4.7),

$$\sum_{v \neq u} \lambda_{v,u}^{\mathsf{UC}} \leq C(\{u\}^c, \{u\}),$$

and

$$\begin{aligned} C(\{u\}^c, \{u\}) &\leq \log\left(1 + (n-1)\textstyle\sum_{v \neq u} |h_{v,u}|^2\right) \\ &\leq \log(1 + (n-1)^2 n^{\alpha}) \\ &\leq K \log(n). \end{aligned}$$

□

The next two lemmas bounds the maximal achievable sum rate across the boundary of the subsquares $V_{\ell,i}(n)$ for $\ell \in \{1, \ldots, \widetilde{L}(n)\}$, and $i \in \{1, \ldots, 4^{\ell}\}$.

Lemma 4.7. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b(n) = O\big(\log^6(n)\big)$ such that for all $V(n) \in \mathcal{V}(n)$, $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$, $\ell \in \{1, \ldots, \widetilde{L}(n)\}$, and*

$i \in \{1, \ldots, 4^\ell\}$, *we have*

$$\sum_{u \in V_{\ell,i}(n)} \sum_{v \notin V_{\ell,i}(n)} \lambda_{u,v}^{\mathsf{UC}} \leq b(n)(4^{-\ell}n)^{2-\min\{3,\alpha\}/2}.$$

Proof. As before, denote by $C(S_1, S_2)$ the MIMO capacity between nodes in S_1 and nodes in S_2, for $S_1, S_2 \subset V$. By the cut-set bound [9, Theorem 14.10.1],

$$\sum_{u \in V_{\ell,i}} \sum_{v \notin V_{\ell,i}} \lambda_{u,v}^{\mathsf{UC}} \leq C(V_{\ell,i}, V_{\ell,i}^c). \tag{4.8}$$

Let

$$\boldsymbol{H}_{S_1,S_2} \triangleq [h_{u,v}]_{u \in S_1, v \in S_2}$$

be the matrix of channel gains between the nodes in S_1 and S_2. Under fast fading

$$C(S_1, S_2) \triangleq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H}) \geq 0: \\ \mathbb{E}(q_{u,u}) \leq P \ \forall u \in S_1}} \mathbb{E}\Big(\log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^\dagger \boldsymbol{Q}(\boldsymbol{H}) \boldsymbol{H}_{S_1,S_2}\right)\Big),$$

and under slow fading

$$C(S, S^c) \triangleq \max_{\substack{\boldsymbol{Q} \geq 0: \\ q_{u,u} \leq P \ \forall u \in S_1}} \log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^\dagger \boldsymbol{Q} \boldsymbol{H}_{S_1,S_2}\right).$$

Denote by $\partial(V_{\ell,i}^c)$ the nodes in $V_{\ell,i}^c$ that are within distance $8(\sqrt{2\log(n)} + 1)$ of the boundary between $A_{\ell,i}^c$ and $A_{\ell,i}$. Using the generalized Hadamard's inequality, yields then that under either fast or slow fading

$$C(V_{\ell,i}, V_{\ell,i}^c) \leq C(V_{\ell,i}, \partial(V_{\ell,i}^c)) + C(V_{\ell,i}, V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)). \tag{4.9}$$

We start by analyzing the first term in the sum in (4.9). Applying Hadamard's inequality again yields

$$C(V_{\ell,i}, \partial(V_{\ell,i}^c)) \leq \sum_{v \in \partial(V_{\ell,i}^c)} C(V_{\ell,i}, \{v\}).$$

Since $V \in \mathcal{V}$, we have

$$|\partial(V_{\ell,i}^c)| \leq 40\log(n)(\sqrt{2\log(n)}+1)(4^{-\ell}n)^{1/2} \leq 120\log^2(n)(4^{-\ell}n)^{1/2}$$

and by the same analysis as in Lemma 4.6, we have

$$C(V_{\ell,i},\{v\}) \leq C(\{v\}^c,\{v\}) \leq \frac{K}{120}\log(n)$$

for some constant K (independent of v). Therefore

$$\begin{aligned} C(V_{\ell,i},\partial(V_{\ell,i}^c)) &\leq 120\log(n)^2(4^{-\ell}n)^{1/2}\frac{K}{120}\log(n) \\ &\leq K\log^3(n)(4^{-\ell}n)^{1/2}. \end{aligned} \tag{4.10}$$

We now analyze the second term in the sum in (4.9). Applying Lemma 3.12,

$$\begin{aligned} &C(V_{\ell,i}, V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)) \\ &\qquad \leq 4\max\Big\{1, \max_{v\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} \sum_{u\in V_{\ell,i}} \frac{r_{u,v}^{-\alpha}}{\sum_{\tilde{v}\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,\tilde{v}}^{-\alpha}}\Big\} \sum_{u\in V_{\ell,i}}\sum_{v\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha}. \end{aligned} \tag{4.11}$$

We now show that the term in parentheses in (4.11) is $O(\log^3(n))$. Fix $u \in V_{\ell,i}$ and denote by d_u the distance of u to the closest point in $V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)$. Note that be definition of $\partial(V_{\ell,i}^c)$, we have

$$d_u \geq 8(\sqrt{2\log(n)}+1).$$

Since, $V \in \mathcal{V}$, there are at least

$$\frac{d_u^2 - 4d_u(\sqrt{2\log(n)}+1)}{2\log(n)} \geq \frac{d_u^2}{4\log(n)}$$

nodes that are at distance at most $3d_u$. Hence

$$\sum_{\tilde{v}\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,\tilde{v}}^{-\alpha} \geq \frac{d_u^{2-\alpha}}{4^{1+\alpha}\log(n)},$$

and therefore for any $v \in V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)$,

$$\begin{aligned}\frac{r_{u,v}^{-\alpha}}{\sum_{\tilde{v}\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,\tilde{v}}^{-\alpha}} &\leq 4^{1+\alpha}\log(n)\frac{r_{u,v}^{-\alpha}}{d_u^{2-\alpha}} \\ &\leq 4^{1+\alpha}\log(n) r_{u,v}^{-2}\Big(\frac{r_{u,v}}{d_u}\Big)^{2-\alpha} \\ &\leq 4^{1+\alpha}\log(n) r_{u,v}^{-2}.\end{aligned} \tag{4.12}$$

Using $V \in \mathcal{V}$, yields that for any $v \in V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)$,

$$\sum_{u\in V_{\ell,i}} \frac{r_{u,v}^{-\alpha}}{\sum_{\tilde{v}\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,\tilde{v}}^{-\alpha}} \leq 4^{1+\alpha}\log(n)\sum_{u\in V_{\ell,i}} r_{u,v}^{-2} \leq 4^{3+\alpha}\ln(2)\log^3(n).$$

Combined with (4.11), this shows that

$$C(V_{\ell,i}, V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)) \leq K'\log^3(n)\sum_{u\in V_{\ell,i}}\sum_{v\in V_{\ell,i}^c} r_{u,v}^{-\alpha} \tag{4.13}$$

for some constant K' (independent of ℓ, and i).

Combining (4.10) and (4.13) with (4.9), we obtain

$$C(V_{\ell,i}, V_{\ell,i}^c) \leq K'\log^3(n)\sum_{u\in V_{\ell,i}}\sum_{v\in V_{\ell,i}^c\setminus\partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} + K\log^3(n)(4^{-\ell}n)^{1/2}. \tag{4.14}$$

Moreover, using again the same arguments as in [38, Theorem 5.2] shows that there exists a constant $\widetilde{K}$ such that for adjacent squares (i.e., sharing a side) $A_{\ell,i}, A_{\ell,j}$,

$$\sum_{u\in V_{\ell,i}}\sum_{v\in V_{\ell,j}\setminus\partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} \leq \widetilde{K}\log^3(n)(4^{-\ell}n)^{2-\min\{3,\alpha\}/2}. \tag{4.15}$$

Consider now two diagonal squares (i.e., sharing a corner point) $A_{\ell,i}, A_{\ell,j}$, and choose

$\tilde{\imath}, \tilde{\jmath}$ such that $A_{\ell,i} \cup A_{\ell,\tilde{\imath}}$ and $A_{\ell,j} \cup A_{\ell,\tilde{\jmath}}$ are adjacent rectangles. Using the same arguments to these rectangles and suitably redefining $\widetilde{K}$ shows that (4.15) holds for diagonal squares as well.

Using this, we now compute the summation in (4.14). Consider "rings" of squares around $A_{\ell,i}$. The first such "ring" contains the (at most) 8 squares neighboring $A_{\ell,i}$. The next "ring" contains at most 16 squares. In general, "ring" k contains at most $8k$ squares. Let

$$\{A_{\ell,j}\}_{j \in I_k}$$

be the squares in "ring" k. Then

$$\sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} = \sum_{k \geq 1} \sum_{j \in I_k} \sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,j} \setminus \partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha}. \tag{4.16}$$

By (4.15),

$$\sum_{j \in I_1} \sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,j} \setminus \partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} \leq 8\widetilde{K} \log^3(n) (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2}. \tag{4.17}$$

Now note that for $k > 1$ and $j \in I_k$, nodes $u \in V_{\ell,i}$ and $v \in V_{\ell,j}$ are at least at distance $r_{u,v} \geq (k-1)(2^{-\ell}\sqrt{n})$. Moreover, since $V \in \mathcal{V}$, each $\{V_{\ell,j}\}_{\ell,j}$ has cardinality at most $4^{-\ell+1} n$. Thus

$$\begin{aligned}
\sum_{k>1} \sum_{j \in I_k} \sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,j} \setminus \partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} &\leq \sum_{k>1} 8k \big(4^{-\ell+1} n\big)^2 \big((k-1)(2^{-\ell}\sqrt{n})\big)^{-\alpha} \\
&= 128 \big(4^{-\ell} n\big)^{2-\alpha/2} \sum_{k>1} k(k-1)^{-\alpha} \\
&= K'' \big(4^{-\ell} n\big)^{2-\alpha/2},
\end{aligned} \tag{4.18}$$

for some $K'' > 0$, and where we have used that $\alpha > 2$. Substituting (4.17) and (4.18) into (4.16) yields

$$\sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,i}^c \setminus \partial(V_{\ell,i}^c)} r_{u,v}^{-\alpha} \leq 8\widetilde{K} \log^3(n) (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2} + K'' \big(4^{-\ell} n\big)^{2-\alpha/2}. \tag{4.19}$$

Combining (4.19) with (4.14) and (4.8) shows that

$$\sum_{u\in V_{\ell,i}}\sum_{v\notin V_{\ell,i}}\lambda_{u,v}^{\mathsf{UC}} \le b(n)(4^{-\ell}n)^{2-\min\{3,\alpha\}/2}.$$

for every $\ell \in \{1,\dots,\widetilde{L}(n)\}, i \in \{1,\dots,4^\ell\}$, and under either fast or slow fading. □

Lemma 4.8. *Under either fast or slow fading, for any $\alpha > 5$, there exists $b(n) = O\big(\log^3(n)\big)$ such that for all $V(n) \in \mathcal{V}(n)$, $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$, $\ell \in \{1,\dots,\widetilde{L}(n)\}$, and $i \in \{1,\dots,4^\ell\}$, we have*

$$\sum_{u\notin V_{\ell,i}(n)}\sum_{v\in V_{\ell,i}(n)}\lambda_{u,v}^{\mathsf{UC}} \le b(n)(4^{-\ell}n)^{1/2}.$$

Proof. By the cut-set bound [9, Theorem 14.10.1],

$$\sum_{u\notin V_{\ell,i}}\sum_{v\in V_{\ell,i}}\lambda_{u,v}^{\mathsf{UC}} \le C(V_{\ell,i}^c, V_{\ell,i}). \tag{4.20}$$

Denote by $\partial V_{\ell,i}$ the nodes in $V_{\ell,i}$ that are within distance one of the boundary between $A_{\ell,i}^c$ and $A_{\ell,i}$. Applying the generalized Hadamard inequality as in Lemma 4.7, we have under either fast or slow fading

$$\begin{aligned} C(V_{\ell,i}^c, V_{\ell,i}) &\le C(V_{\ell,i}^c, \partial V_{\ell,i}) + C(V_{\ell,i}^c, V_{\ell,i}\setminus\partial V_{\ell,i}) \\ &\le K\log^3(n)(4^{-\ell}n)^{1/2} + C(V_{\ell,i}^c, V_{\ell,i}\setminus\partial V_{\ell,i}), \end{aligned} \tag{4.21}$$

for some constant K.

For the second term in (4.21), we have the following upper bound from Theorem 2.1 in [21]:

$$C(V_{\ell,i}^c, V_{\ell,i}\setminus\partial V_{\ell,i}) \le \sum_{v\in V_{\ell,i}\setminus\partial V_{\ell,i}}\Big(\sum_{u\in V_{\ell,i}^c} r_{u,v}^{-\alpha/2}\Big)^2.$$

Now, consider $v \in V_{\ell,i}\setminus\partial V_{\ell,i}$ and let d_v be the distance of v to the closest node in

$V_{\ell,i}^c$. Then, using $V \in \mathcal{V}$,

$$\sum_{u \in V_{\ell,i}^c} r_{u,v}^{-\alpha/2} \leq \widetilde{K} \log(n) d_v^{2-\alpha/2},$$

for some constant $\widetilde{K}$, and hence for $\alpha > 5$,

$$\begin{aligned} C(V_{\ell,i}^c, V_{\ell,i} \setminus \partial V_{\ell,i}) &\leq \sum_{v \in V_{\ell,i} \setminus \partial V_{\ell,i}} \widetilde{K}^2 \log^2(n) d_v^{4-\alpha} \\ &\leq K' \log^3(n) (4^{-\ell} n)^{1/2}. \end{aligned}$$

Combined with (4.21), this proves Lemma 4.8 for traffic from $V_{\ell,i}^c$ to $V_{\ell,i}$. □

4.4.3 Achievability Lemmas

In this section, we prove auxiliary achievability results. Recall that a permutation traffic is a traffic pattern in which each node is source and destination exactly once. Call the corresponding source-destination pairing $\Pi \subset V(n) \times V(n)$ a *permutation pairing*. The lemma below analyzes the performance achievable with either hierarchical relaying (for $\alpha \in (2, 3]$) or multi-hop communication (for $\alpha > 3$) applied simultaneously to transmit permutation traffic in several disjoint regions in the network.

Lemma 4.9. *Under fast fading, for any $\alpha > 2$, there exists $b(n) \geq n^{-o(1)}$ such that for all $V(n) \in \mathcal{V}(n)$, $\ell \in \{0, \ldots, \widetilde{L}(n)\}$, $i \in \{1, \ldots, 4^\ell\}$, and all permutation source-destination pairings Π_i on $V_{\ell,i}(n)$, there exists $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$ such that*

$$\min_{i \in \{1,\ldots,4^\ell\}} \min_{(u,v) \in \Pi_i} \lambda_{u,v}^{\mathsf{UC}} \geq b(n) (4^{-\ell} n)^{1 - \min\{3,\alpha\}/2}.$$

The same statement holds with probability $1 - o(1)$ as $n \to \infty$ in the slow fading case.

Consider the source-destination pairing $\Pi \triangleq \cup_i \Pi_i$ with Π_i as in Lemma 4.9. This is a permutation pairing, since each Π_i is a permutation pairing on $V_{\ell,i}(n)$ and since the $\{V_{\ell,i}(n)\}_i$ are disjoint. Lemma 4.9 states that all source-destination pairs in Π

can communicate at a per-node rate of at least $n^{-o(1)}(4^{-\ell}n)^{1-\min\{3,\alpha\}/2}$. Note that, due to the locality of the traffic pattern, this can be considerably better than the $n^{1-\min\{3,\alpha\}/2-o(1)}$ per-node rate achieved by standard hierarchical relaying or multi-hop communication.

Proof. We shall use either hierarchical relaying (for $\alpha \in (2,3]$) or multi-hop (for $\alpha > 3$) to communicate within each square $V_{\ell,i}$. We operate every fourth of the $V_{\ell,i}$ simultaneously, and show that the added interference due to this spatial re-use results only in a constant factor loss in rate.

Consider first $\alpha \in (2,3]$ and fast fading. The squares $A_{\ell,i}$ at level ℓ have an area of

$$a_\ell \triangleq 4^{-\ell}n.$$

We will use hierarchical relaying within each of the $\{A_{\ell,i}\}_i$. Applying Theorem 3.1) with $\delta = 1/4$, it is sufficient to show that we can partition each $A_{\ell,i}$ into

$$a_\ell^{\frac{1}{1+\log^{-1/4}(a_\ell)}}$$

subsquares, each of which contains a number of nodes proportional to the area (see Section 3.10.4). In other words, we partition A into subsquares of size

$$\begin{aligned} a_\ell^{1-\frac{1}{1+\log^{-1/4}(a_\ell)}} &\geq a_{\widetilde{L}(n)}^{\frac{\log^{-1/4}(n)}{1+\log^{-1/4}(n)}} \\ &\geq a_{\widetilde{L}(n)}^{\log^{-1/3}(n)} \\ &= n^{\log^{-5/6}(n)} \\ &\geq n4^{-\log(n)(1-\log^{-5/6}(n))}. \end{aligned}$$

Since $V \in \mathcal{V}$, all these subsquares contain a number of nodes proportional to their area, and hence this shows that all

$$\{A_{i,\ell}\}_{\ell\in\{0,\ldots,\widetilde{L}(n)\},i\in\{1,\ldots,4^\ell\}}$$

are simultaneously regular enough for hierarchical relaying to be successful under fast fading. By Theorem 3.1, this achieves a per-node rate of

$$\lambda_{u,v}^{\mathsf{UC}} \geq n^{-o(1)}(4^{-\ell}n)^{1-\alpha/2} \tag{4.22}$$

for any $(u, v) \in \Pi_i$.

We now show that (4.22) holds with high probability also under slow fading. By Theorem 3.1, for $V \in \mathcal{V}$ hierarchical relaying is successful under slow fading for all permutation traffic on V with probability at least

$$1 - \exp\Big(- 2^{K \log^{3/4}(n)}\Big)$$

for some constant K. Hence, hierarchical relaying is successful for all permutation traffic on $V_{\ell,i}$ with probability at least

$$\begin{aligned} 1 - \exp\Big(- 2^{K \log^{3/4}(4^{-\ell}n)}\Big) &\geq 1 - \exp\Big(- 2^{K \log^{3/4}(4^{-\widetilde{L}(n)}n)}\Big) \\ &= 1 - \exp\Big(- 2^{K \log^{3/8}(n)}\Big). \end{aligned}$$

And hence hierarchical relaying is successful under slow fading for all $\ell \in \{1, \ldots, \widetilde{L}(n)\}$ and all permutation traffic on every $\{V_{\ell,i}\}_{i=1}^{4^\ell}$ with probability at least

$$\begin{aligned} 1 - \widetilde{L}(n)4^{\widetilde{L}(n)} \exp\Big(- 2^{K \log^{3/8}(n)}\Big) &\geq 1 - n^2 \exp\Big(- 2^{K \log^{3/8}(n)}\Big) \\ &\geq 1 - o(1) \end{aligned}$$

as $n \to \infty$.

We now argue that the additional interference from spatial re-use results only in a constant loss in rate. This follows from the same arguments as in the proof of Theorem 3.1 (with the appropriate modifications for slow fading as described there). Intuitively, this is the case since the interference from a square at distance r is attenuated by a factor $r^{-\alpha}$, which, since $\alpha > 2$, is summable. Hence the combined interference has power on the order of the receiver noise, resulting in only a constant factor loss in

rate.

For $\alpha > 3$, the argument is similar — instead of hierarchical relaying we now use multi-hop communication. For $V \in \mathcal{V}$ and under either fast or slow fading, this achieves a per-node rate of

$$\lambda_{u,v}^{\mathsf{UC}} \geq n^{-o(1)}(4^{-\ell}n)^{-1/2} \tag{4.23}$$

for any $(u, v) \in \Pi_i$. Combining (4.22) and (4.23) yields the desired result. □

4.5 Proof of Theorem 4.1

The proof of Theorem 4.1 relies on the construction of a capacitated (noiseless, wireline) graph G and linking its performance under routing to the performance of the wireless network. This graph $G = (V_G, E_G)$ is constructed as follows. G is a full tree (i.e., all its leaf nodes are on the same level). G has n leaves, each of them representing an element of $V(n)$. To simplify notation, we assume that $V(n) \subset V_G$, so that the leaves of G are exactly the elements of $V(n) \subset V_G$. Whenever the distinction is relevant, we use u, v for nodes in $V(n) \subset V_G$ and μ, ν for nodes in $V_G \setminus V(n)$ in the following. The internal nodes of G correspond to $V_{\ell,i}(n)$ for all $\ell \in \{0, \ldots, \widetilde{L}(n)\}$, $i \in \{1, \ldots, 4^\ell\}$, with hierarchy induced by the one on $A(n)$. In particular, let μ and ν be internal nodes in V_G and let $V_{\ell,i}(n)$ and $V_{\ell+1,j}(n)$ be the corresponding subsets of $V(n)$. Then ν is a child node of μ if $V_{\ell+1,j}(n) \subset V_{\ell,i}(n)$.

In the following, we will assume $V(n) \in \mathcal{V}(n)$, which holds with probability $1-o(1)$ as $n \to \infty$ by Lemma 4.5. With this assumption, nodes in V_G at level $\ell < \widetilde{L}(n)$ have 4 children each, nodes in V_G at level $\ell = \widetilde{L}(n)$ have between $4^{-\widetilde{L}(n)-1}n$ and $4^{-\widetilde{L}(n)+1}n$ children, and nodes in V_G at level $\ell = \widetilde{L}(n) + 1$ are the leaves of the tree (see Figure 4-5 below and Figure 4-3 in Section 4.3).

For $\mu \in V_G$, denote by $\mathcal{L}(\mu)$ the leaf nodes of the subtree of G rooted at μ. Note that, by construction of the graph G, $\mathcal{L}(\mu) = V_{\ell,i}(n)$ for some ℓ and i. To understand the relation between V_G and $V(n)$, we define the *representative* $\mathcal{R} : V_G \to 2^{V(n)}$ of μ

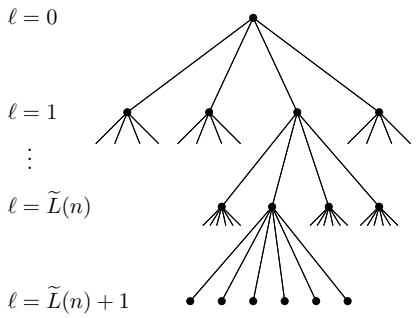

Figure 4-5: Communication graph G constructed in the proof of Theorem 4.1. Nodes on levels $\ell \in \{0, \ldots, \widetilde{L}(n) - 1\}$ have four children each, nodes on level $\ell = \widetilde{L}(n)$ have $\Theta\big(n^{\log^{-1/2}(n)}\big)$ children each. The total number of leaf nodes is n, one representing each node in the wireless network $V(n)$. An internal node in G at level $\ell \in \{0, \ldots, \widetilde{L}(n)\}$ represents the collection of nodes in $V_{\ell,i}(n)$ for some i.

as follows. For a leaf node $u \in V(n) \subset V_G$ of G, let

$$\mathcal{R}(u) \triangleq \{u\}.$$

For $\mu \in V_G$ at level $\widetilde{L}(n)$, choose $\mathcal{R}(\mu) \subset \mathcal{L}(\mu) \subset V(n)$ such that

$$|\mathcal{R}(\mu)| = 4^{-\widetilde{L}(n)-1} n.$$

This is possible since $V(n) \in \mathcal{V}(n)$ by assumption. Finally, for $\mu \in V_G$ at level $\ell < \widetilde{L}(n)$, and with children $\{\nu_i\}_{j=1}^4$, let

$$\mathcal{R}(\mu) \triangleq \bigcup_{j=1}^{4} \mathcal{R}(\nu_j).$$

We now define an edge capacity $c_{\mu,\nu}$ for each edge $(\mu, \nu) \in E_G$. If μ is a leaf of G and ν its parent, set

$$c_{\mu,\nu} = c_{\nu,\mu} \triangleq 1. \tag{4.24}$$

If μ is an internal node at level ℓ in G and ν its parent, then set

$$c_{\mu,\nu} = c_{\nu,\mu} \triangleq (4^{-\ell} n)^{2-\min\{3,\alpha\}/2}. \tag{4.25}$$

Having chosen edge capacities on G, we can now define the set $\Lambda_G^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n \times n}$ of feasible unicast traffic matrices between leaf nodes of G. In other words, $\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}(n)$ if messages at the leaf nodes of G can be routed to their destinations (which are also leaf nodes) over G at rates λ^{UC} while respecting the capacity constraints on the edges of G.

We first prove the achievability part of Theorem 4.1. The next lemma shows that if traffic can be routed over the tree G, then approximately the same traffic can be transmitted reliably over the wireless network.

Lemma 4.10. *Under fast fading, for any $\alpha > 2$, there exists $b(n) \geq n^{-o(1)}$ such that for any $V(n) \in \mathcal{V}(n)$,*

$$b(n)\Lambda_G^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n).$$

The same statement holds for slow fading with probability $1 - o(1)$ as $n \to \infty$.

Proof. Assume $\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}$, i.e., traffic can be routed between the leaf nodes of G at a rate λ^{UC}, we need to show that $n^{-o(1)}\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ (i.e., almost the same flow can be reliably transmitted over the wireless network). We use the three-layer communication architecture introduced in Section 4.3 to establish this result.

Recall the three layers of this architecture: the routing, cooperation, and physical layers. The layers of this communication scheme operate as follows. In the routing layer, we treat the wireless network as the graph G and route the messages over the edges of G. The cooperation layer provides this tree abstraction to the routing layer by distributing and concentrating messages over subsets of the wireless network. The physical layer implements this distribution and concentration of messages by dealing with interference and noise.

Consider first the routing layer, and assume that the tree abstraction G can be implemented in the wireless network with only a $n^{-o(1)}$ factor loss. Since $\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}$ by assumption, we then know that the routing layer will be able to reliably transmit messages at rates $n^{-o(1)}\lambda^{\mathsf{UC}}$. We now show that the tree abstraction can indeed be implemented in the wireless network.

This tree abstraction is provided to the routing layer by the cooperation layer. We will show that the operation of the cooperation layer satisfies the following invariance property: If a message is located at a node $\mu \in G$ in the routing layer, then the same message is evenly distributed over all nodes in $\mathcal{R}(\mu)$ in the wireless network. In other words, all nodes $u \in \mathcal{R}(\mu) \subset V$ have access to a distinct part of length $1/|\mathcal{R}(\mu)|$ of the message.

Consider first a leaf node $u \in V \subset V_G$ in G, and assume the routing layer calls upon the cooperation layer to send a message to its parent $\nu \in V_G$ in G. Note first that u is also an element of V, and it has access to the entire message to be sent over G. Since for leaf nodes $\mathcal{R}(u) = \{u\}$, this shows that the invariance property is satisfied at u. The message is split at u into $|\mathcal{R}(\nu)|$ parts of equal length, and one part is sent to each node in $\mathcal{R}(\nu)$ over the wireless network. In other words, we distribute the message over the wireless network by a factor of $|\mathcal{R}(\nu)|$. Hence the invariance property is also satisfied at μ.

Consider now an internal node $\mu \in V_G$, and assume the routing layer calls upon the cooperation layer to send a message to its parent node $\nu \in V_G$. Note that since all traffic in G originates at the leaf nodes of G (which are the actual nodes in the wireless network), a message at μ had to traverse all levels below μ in the tree G. We assume that the invariance property holds up to level μ in the tree, and show that it is then also satisfied at level ν. By the induction hypothesis, each node $u \in \mathcal{R}(\mu)$ has access to a distinct part of length $1/|\mathcal{R}(\mu)|$ of the message. Each such node u splits its message part into four distinct parts of equal length. Node u keeps one part for itself, and sends the other three parts to nodes in $\mathcal{R}(\nu)$. Since $|\mathcal{R}(\nu)| = 4|\mathcal{R}(\mu)|$, this can be performed such that each node in $\mathcal{R}(\nu)$ obtains exactly one message part. In other words, we distribute the message by a factor four over the wireless network, and the invariance property is satisfied at $\nu \in V_G$.

Operation along edges down the tree (i.e., towards the leaf nodes) is similar, but instead of distributing messages, we now concentrate them over the wireless network. To route a message from a node $\mu \in V_G$ with internal children $\{\nu_j\}_{j=1}^4$ to one of them (say ν_1) in the routing layer, the cooperation layer sends the message parts from each

$\{\mathcal{R}(\nu_j)\}_{j=2}^4$ to a corresponding node in $\mathcal{R}(\nu_1)$ and combines them there. In other words, we concentrate the message by a factor four over the wireless network.

To route a message to a leaf node $u \in V \subset V_G$ from its parent ν in G in the routing layer, the cooperation layer sends the corresponding message parts at each node $\mathcal{R}(\nu)$ to u over the wireless network. Thus, again we concentrate the message over the network, but this time by a factor of $|\mathcal{R}(\nu)|$. Both these operations along edges down the tree preserve the invariance property. This shows that the invariance property is preserved by all operations induced by the routing layer in the cooperation layer.

Finally, to actually implement this distribution and concentration of messages, the cooperation layer calls upon the physical layer. Note that at the routing layer, all edges of the tree can be routed over simultaneously. Therefore, the cooperation layer can potentially call the physical layer to perform distribution and concentration of messages over all sets $\{\mathcal{R}(\mu)\}_{\mu \in V_G}$ simultaneously. The function of the physical layer is to schedule all these operations and to deal with the resulting interference as well as with channel noise.

This scheduling is done as follows. First, the physical layer time shares between communication up the tree and communication down the tree (i.e., between distribution and concentration of messages). This results in a loss of a factor $1/2$ in rate. The physical layer further time shares between all the $\widetilde{L}(n) + 1$ internal levels of the tree, resulting in a further $\frac{1}{\widetilde{L}(n)+1}$ factor loss in rate. Hence, the total rate loss by this time sharing is

$$\frac{1}{2(\widetilde{L}(n) + 1)}. \tag{4.26}$$

Consider now the operations within some level $\ell \in 1, \ldots, \widetilde{L}(n)$ in the tree (i.e., for edge (μ, ν) on this level, neither μ nor ν is a leaf node). We show that the rate at which the physical layer implements the edge (μ, ν) is equal to $n^{-o(1)} c_{\mu,\nu}$, i.e., only a $n^{o(1)}$ factor less than the capacity of the edge (μ, ν) in the tree G. Note first that the distribution or concentration of traffic induced by the cooperation layer to implement one edge e at level ℓ is restricted to $V_{\ell-1,i}$ for some $i = i(e)$. We can thus partition

the edges at level ℓ into $\{E_G^j\}_{j=1}^4$ such that the four sets

$$\bigcup_{e \in E_G^j} V_{\ell-1,i(e)}$$

of nodes are disjoint. Time sharing between these four sets yields an additional loss of a factor 1/4 in rate. Fix one such value of j, and consider the operations induced by the cooperation layer in the set corresponding to j. We consider communication up the tree (i.e., distribution of messages), the analysis for communication down the tree is similar. For a particular edge $(\mu, \nu) \in E_G^j$ with ν the parent of μ, each node $u \in \mathcal{R}(\mu)$ has split its message part into four parts, three of which need to be sent to other nodes in $\mathcal{R}(\nu)$. Moreover, this assignment of destination nodes in $\mathcal{R}(\nu)$ to u is performed such that no node in $\mathcal{R}(\nu)$ is destination more than once. In other words, each node in $\mathcal{R}(\mu)$ is source exactly three times and each node in $\mathcal{R}(\nu)$ is destination at most once. This can be written as three source-destination pairings $\{\Pi_{i(\mu,\nu)}^k\}_{k=1}^3$, on $V_{\ell-1,i(\mu,\nu)}$. Moreover, each such $\Pi_{i(\mu,\nu)}^k$ can be understood as a subset of a permutation source-destination pairing. We time share between the three values of k (yielding an additional loss of a factor 1/3 in rate). Now, for each value of k, Lemma 4.9 shows that by using either hierarchical relaying (for $\alpha \in (2, 3]$) or multi-hop communication for ($\alpha > 3$), we can communicate according to $\{\Pi_{i(e)}^k\}_{e \in E_G^j}$ at a per-node rate of

$$n^{-o(1)}(4^{-\ell-1}n)^{1-\min\{3,\alpha\}/2}$$

under fast fading, and with probability $1 - o(1)$ also under slow fading.[3] Since $\mathcal{R}(\mu)$ contains $4^{-\ell-1}n$ nodes, and accounting for the loss (4.26) for time sharing between the levels in G and the additional loss of factors 1/4 and 1/3 for time sharing between j and k, the physical layer implements an edge capacity for e at level ℓ of

$$\frac{1}{2(\widetilde{L}(n)+1)} \cdot \frac{1}{4} \cdot \frac{1}{3} \cdot 4^{-\ell-1}n \cdot n^{-o(1)}(4^{-\ell-1}n)^{1-\min\{3,\alpha\}/2}$$

[3]Note that Lemma 4.9 actually shows that all permutation traffic for every value of ℓ can be transmitted with high probability under slow fading. In other words, with high probability all levels of G can be implemented successfully under slow fading.

$$= n^{-o(1)}(4^{-\ell}n)^{2-\min\{3,\alpha\}/2} = n^{-o(1)}c_e.$$

Consider now the operations within level $\ell = \widetilde{L}(n) + 1$ in the tree (i.e., for edge (u, ν) on this level, u is a leaf node). We show that the rate at which the physical layer implements the edge (u, ν) is equal to $n^{-o(1)}c_{u,\nu}$. We again consider only communication up the tree (i.e., distribution of messages in the cooperation layer), communication down the tree is performed in a similar manner. The traffic induced by the cooperation layer at level $\widetilde{L}(n)+1$ is within the sets $V_{\widetilde{L}(n),i}$ for $i = \{1, \ldots, 4^{\widetilde{L}(n)}\}$. Consider now communication within one $V_{\widetilde{L}(n),i}$, and assume without loss of generality that in the routing layer every node $u \in V_{\widetilde{L}(n),i}$ needs to send traffic along the edge (u, ν). In the physical layer, we need to distribute a $1/|\mathcal{R}(\nu)|$ fraction of this traffic from each node $u \in V_{\widetilde{L}(n),i}$ to every node in $\mathcal{R}(\nu) \subset V_{\widetilde{L}(n),i}$. This can be expressed as $|V_{\widetilde{L}(n),i}|$ source-destination pairings, and we time share between them. Accounting for the fact that only $1/|\mathcal{R}(\nu)|$ of traffic needs to be sent according to each pairing and since $V \in \mathcal{V}$, this results in a time sharing loss of at most a factor

$$\frac{|\mathcal{R}(\nu)|}{|V_{\widetilde{L}(n),i}|} \leq \frac{1}{16}.$$

Now, using Lemma 4.9, all these source-destination pairings in all subsquares $\{V_{\widetilde{L}(n),i}\}$ can be implemented simultaneously at a per node rate of

$$n^{-o(1)}(4^{-\widetilde{L}}n)^{1-\min\{3,\alpha\}/2} \geq n^{-o(1)}(n^{\log^{-1/2}(n)})^{-1/2} \geq n^{-o(1)}.$$

Accounting for the loss (4.26) for time sharing between the levels in G, the additional factor $1/16$ loss for time sharing within each $V_{\widetilde{L}(n),i}$, the physical layer implements an edge capacity for e at level $\ell = \widetilde{L}(n) + 1$ of

$$\frac{1}{2(\widetilde{L}(n)+1)} \cdot \frac{1}{16} \cdot n^{-o(1)} = n^{-o(1)} = n^{-o(1)}c_e,$$

under either fast or slow fading.

Together, this shows that the physical and cooperation layers provide the tree

abstraction G to the routing layer with edge capacities of only a factor $n^{-o(1)}$ loss. Hence, if messages can be routed at rates λ^{UC} between the leaf nodes of G, then messages can be reliably transmitted over the wireless network at rates $n^{-o(1)}\lambda^{\mathsf{UC}}$. Hence

$$\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}} \;\Rightarrow\; n^{-o(1)}\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}},$$

and noting that the $n^{-o(1)}$ factor is uniform in λ^{UC}, this shows that

$$n^{-o(1)}\Lambda_G^{\mathsf{UC}} \subset \Lambda^{\mathsf{UC}}.$$

□

We have seen that the unicast capacity region $\Lambda_G^{\mathsf{UC}}(n)$ of the graph G under routing is (appropriately scaled) an inner bound to the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network. The next lemma shows that $\Lambda_G^{\mathsf{UC}}(n)$ is equal to the approximate unicast capacity region $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ of the wireless network as defined in (4.2). Combining Lemmas 4.5, 4.10, and 4.11 below, yields that with high probability

$$n^{-o(1)}\widehat{\Lambda}^{\mathsf{UC}}(n) \subset n^{-o(1)}\Lambda_G^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n),$$

proving the achievability part of Theorem 4.1.

Lemma 4.11. *For any $\alpha > 2$ and any $V(n) \in \mathcal{V}(n)$,*

$$\widehat{\Lambda}_1^{\mathsf{UC}}(n) = \Lambda_G^{\mathsf{UC}}(n).$$

Proof. We first relate the total traffic across an edge e in the graph G to the total traffic across a cut $V_{\ell,i}$ for some ℓ and i.

Consider an edge $e = (\mu, \nu) \in E_G$, and assume first that e connects a node at level $\ell + 1$ and ℓ in the tree with $\ell < \widetilde{L}(n)$. We slight abuse of notation, set

$$c_e \triangleq c_{\mu,\nu}.$$

Note first that by (4.25) we have

$$c_e = (4^{-\ell} n)^{2-\min\{3,\alpha\}/2}. \tag{4.27}$$

Moreover, since G is a tree, removing the edge e from E_G separates the tree into two connected components, say $S_1, S_2 \subset V_G$. Consider now the leaf nodes in S_1. By the construction of the tree structure of G, these leaf nodes are either equal to $V_{\ell,i}$ or $V_{\ell,i}^c$ for some $i \in \{1, \ldots, 4^\ell\}$. Assume without loss of generality that they are equal to $V_{\ell,i}$. Then $V_{\ell,i}^c$ are the leaf nodes in S_2. Now since traffic is only assumed to be between leaf nodes of G, the total traffic demand between S_1 and S_2 is equal to

$$\sum_{u \in V_{\ell,i}} \sum_{w \in V_{\ell,i}^c} (\lambda_{u,w}^{\mathsf{UC}} + \lambda_{w,u}^{\mathsf{UC}}). \tag{4.28}$$

By the tree structure of G, all this traffic has to be routed over edge e.

Consider now an edge e connecting a node at level $\widetilde{L}(n) + 1$ and $\widetilde{L}(n)$, i.e., a leaf node u to its parent ν. Then, by (4.24),

$$c_e = 1, \tag{4.29}$$

and the total traffic crossing the edge e is equal to

$$\sum_{w \neq u} (\lambda_{u,w}^{\mathsf{UC}} + \lambda_{w,u}^{\mathsf{UC}}). \tag{4.30}$$

We now show that

$$\widehat{\Lambda}_1^{\mathsf{UC}} \subset \Lambda_G^{\mathsf{UC}}. \tag{4.31}$$

Assume $\lambda^{\mathsf{UC}} \in \widehat{\Lambda}_1^{\mathsf{UC}}$, then

$$\sum_{u \in V_{\ell,i}} \sum_{w \in V_{\ell,i}^c} (\lambda_{u,w}^{\mathsf{UC}} + \lambda_{w,u}^{\mathsf{UC}}) \leq (4^{-\ell} n)^{2-\min\{3,\alpha\}}$$

for all $\ell \in \{1, \ldots, \widetilde{L}(n)\}, i \in \{1, \ldots, 4^\ell\}$, and

$$\sum_{w \neq u} (\lambda_{u,w}^{\mathsf{UC}} + \lambda_{w,u}^{\mathsf{UC}}) \leq 1$$

for all $u \in V$. By (4.27), (4.28), (4.29), (4.30), this implies that the traffic demand across each edge e of the graph G is less than its capacity c_e. Since G is a tree, this implies that λ^{UC} can be routed over G, i.e., $\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}$. This proves (4.31).

We now show that

$$\Lambda_G^{\mathsf{UC}} \subset \widehat{\Lambda}_1^{\mathsf{UC}}. \tag{4.32}$$

Assume $\lambda^{\mathsf{UC}} \in \widehat{\Lambda}_G^{\mathsf{UC}}$. This implies λ^{UC} can be routed over G, and hence the total flow across each edge is less than the capacity of that edge. By (4.27), (4.28), (4.29), (4.30), this implies $\lambda^{\mathsf{UC}} \in \widehat{\Lambda}_1^{\mathsf{UC}}$, from which (4.32) follows. □

We now turn to the converse part of Theorem 4.1. The next lemma shows that $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ (appropriately scaled) is an outer bound to the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network. Combined with Lemma 4.5, this shows that with high probability

$$\Lambda^{\mathsf{UC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_2^{\mathsf{UC}}(n),$$

proving the converse part of Theorem 4.1.

Lemma 4.12. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b(n) = O(\log^6(n))$ such that for any $V(n) \in \mathcal{V}(n)$,*

$$\Lambda^{\mathsf{UC}}(n) \subset b(n)\widehat{\Lambda}^{\mathsf{UC}}(n).$$

Proof. Assume $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$. By Lemma 4.7, we have for any $\ell \in \{1, \ldots, \widetilde{L}(n)\}$ and $i \in \{1, \ldots, 4^\ell\}$,

$$\sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,i}^c} \lambda_{u,v}^{\mathsf{UC}} \leq K \log^6(n)(4^{-\ell}n)^{2-\min\{3,\alpha\}/2}, \tag{4.33}$$

for some constant K not depending on λ^{UC}. Moreover, Lemma 4.6 shows that for all

$u \in V$

$$\sum_{v \neq u} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}) \leq 2\widetilde{K} \log(n). \tag{4.34}$$

Combining (4.33) and (4.34) proves that there exists $b(n) = O(\log^6(n))$ such that $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ implies $\lambda^{\mathsf{UC}} \in b(n)\widehat{\Lambda}_2^{\mathsf{UC}}$, proving the lemma. □

4.6 Proof of Theorem 4.2

Assume $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$. By Lemmas 4.7 and 4.8, we have for any $\ell \in \{1, \ldots, \widetilde{L}(n)\}$, $i \in \{1, \ldots, 4^\ell\}$, and $\alpha > 5$,

$$\max \left\{ \textstyle\sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,i}^c} \lambda_{u,v}^{\mathsf{UC}}, \sum_{u \in V_{\ell,i}^c} \sum_{v \in V_{\ell,i}} \lambda_{u,v}^{\mathsf{UC}}, \right\} \leq K \log^6(n) (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2},$$

for some constant K not depending on λ^{UC}. Hence

$$\sum_{u \in V_{\ell,i}} \sum_{v \in V_{\ell,i}^c} (\lambda_{u,v}^{\mathsf{UC}}, + \lambda_{v,u}^{\mathsf{UC}}) \leq 2K \log^6(n) (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2}.$$

Combined with the outer bound in Theorem 4.1, this shows that there exists $b(n) = O(\log^6(n))$ such that for $\alpha > 5$, $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ implies $\lambda^{\mathsf{UC}} \in b(n)\widehat{\Lambda}_1^{\mathsf{UC}}$, proving Theorem 4.2.

4.7 Discussion

We discuss several aspects and extensions of the three-layer architecture introduced in Section 4.3 and used to show achievability in Theorem 4.1. In Section 4.7.1, we comment on the various tree structures used in the three-layer architecture. In Section 4.7.2, we show that for certain values of α the bounds in Theorem 4.1 can be significantly sharpened. In Section 4.7.3, we point out how the results discussed so far can be used to obtain the scaling of the unicast capacity region of dense networks (where n nodes are randomly placed on a square of unit area). Section 4.7.4 contains design guidelines that can be obtained from the scaling results for the unicast capacity

region presented in this chapter.

4.7.1 Tree Structures

There are two distinct tree structures that are used in the construction of the three-layer communication scheme proposed in this chapter — one explicit and one implicit. These two tree structures appear in different layers of the communication scheme and serve different purposes.

The first (explicit) tree structure is given by the tree G utilized in the routing layer and implemented in the cooperation layer. The main purpose of this tree structure is to perform localized load balancing. In fact, the distribution and concentration of traffic is used to avoid unnecessary bottlenecks. Note that the tree G is used by the scheme for any value of α.

The second (implicit) tree structure occurs in the physical layer. This tree structure appears only for values of $\alpha \in (2, 3]$, when the physical layer operates using the hierarchical relaying scheme (see Chapter 3). It is this hierarchical structure of this scheme that can equivalently be understood as a tree. The purpose of this second tree structure is to enable multiple antenna communication, i.e., to perform cooperative communication.

4.7.2 Second-Order Asymptotics

The scaling result in Theorems 4.1 and 4.2 are up to a factor $n^{\pm o(1)}$ and hence preserve information at scale n^{β} for constant β (see also the discussion in Section 4.1.2). Here we take a closer look at the behavior of this $n^{\pm o(1)}$ factor, and show that in certain situations it can be significantly sharpened.

Note first that the outer bound in Theorems 4.1 and 4.2 hold up to a factor $\log^6(n)$, i.e., poly-logarithmic in n. However, the inner bound holds only up to the aforementioned $n^{-o(1)}$ factor. A closer look at the proof of the theorems reveals that

the precise inner bound is of order

$$n^{-O\left(\log^{-1/4}(n)\right)},$$

and with a more careful analysis (see Chapter 3 for the details), this can be sharpened to essentially

$$n^{-O\left(\log^{-1/2}(n)\right)}.$$

The exponent $\log^{-1/2}(n)$ in the inner bound has two causes. The first is the use of hierarchical relaying (for $\alpha \in (2,3]$). The second is the operation of the physical layer at level $\widetilde{L}(n)+1$ of the tree (i.e., to implement communication between the leaf nodes of G and their parents). Indeed at that level, we are operating on a square of area

$$4^{-\widetilde{L}(n)}n = n^{\log^{1/2}(n)},$$

and the loss is essentially inversely proportional to that area. Now, the reason why $\widetilde{L}(n)$ cannot be chosen to be larger (to make this loss smaller), is because hierarchical relaying requires a certain amount of regularity in the node placement, which can only be guaranteed for large enough areas.

This suggests that for the $\alpha > 3$ regime, where multi-hop communication is used at the physical layer instead of hierarchical relaying, we might be able to significantly improve the inner bound. To this end, we have to choose more levels in the tree G, such that at the last level before the tree nodes, we are operating on a square that has an area of order $\log(n)$. Changing the three-layer architecture in this manner, for $\alpha > 3$ the inner bound can be improved to be poly-logarithmic in n as well. Combined with the poly-logarithmic outer bound, this yields a poly-logarithmic approximation for $n^2 - n$ out of n^2 total dimensions of the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ for $\alpha \in (3,5]$, and a poly-logarithmic approximation for the entire unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ for $\alpha > 5$.

4.7.3 Dense Networks

So far, we have only discussed *extended* networks, i.e., n nodes are located on a square of area n. We now briefly sketch how these results can be recast for *dense* networks, in which n nodes are located on a square of unit area. It suffices to notice that by rescaling power by a factor $n^{-\alpha/2}$, a dense network can essentially be transformed into an equivalent extended network with path-loss exponent α (see also [38]). Hence the scaling of the unicast capacity region for dense networks can be obtained from the scaling result for extended networks by taking a limit as $\alpha \to 2$.

The resulting approximate unicast capacity region $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ has a particularly simple shape in this limit. In fact, the only constraints in (4.2) that can be tight are at level $\ell = \log(n)$. This results in the following approximate unicast capacity region for dense networks:

$$\widehat{\Lambda}_1^{\mathsf{UC}}(n) \triangleq \left\{ \lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n} : \sum_{v \neq u} (\lambda_{u,v}^{\mathsf{UC}} + \lambda_{v,u}^{\mathsf{UC}}) \leq 1, \forall\ u \in V(n) \right\},$$

and we obtain that for dense networks, for any $\alpha > 2$,

$$n^{-o(1)} \widehat{\Lambda}_1^{\mathsf{UC}}(n) \subset \Lambda^{\mathsf{UC}}(n) \subset O(\log^6(n)) \widehat{\Lambda}_1^{\mathsf{UC}}(n).$$

Note that, in contrast to the extended case, this results in an approximate characterization of the entire unicast region for all $\alpha > 2$.

4.7.4 Design Guidelines

The results presented in this chapter suggest the following design guidelines for the construction and operation of large wireless networks. First, it shows that load balancing is crucial for optimal operation. This is implemented in the proposed three-layer architecture by spreading traffic over clusters of neighbors, but could be achieved in other ways as well. A communication protocol not using such load balancing could create bottlenecks that prevent it from operating optimally. It is worth pointing out that in the case of random source-destination pairing with uniform rate no such load

balancing is necessary since the traffic demand itself is already balanced.

Second, the load balancing has to be done in a localized manner. In the three-layer architecture, this is achieved by choosing the cluster over which the load balancing is performed to be those nodes that are in the same smallest square $V_{\ell,i}(n)$ that contains both the source and its destination. Again, other ways to perform this load balancing are possible, however for optimal operation this load balancing must have the same localized structure.

Finally, recall that for random node placement and random source-destination pairing with uniform rate, hierarchical cooperation is order optimal for $\alpha \in (2, 3]$, and multi-hop communication is order optimal for $\alpha > 3$. The same threshold phenomenon occurs also for random node placement with heterogeneous traffic. In the three-layer architecture, this is visible in the physical layer, which uses hierarchical relaying for $\alpha \in (2, 3]$ and multi-hop communication for $\alpha > 3$. In other words, for optimal transmission of heterogeneous traffic over randomly placed nodes, global cooperation is necessary for small path-loss exponents, and local cooperation is sufficient for large path-loss exponents. Note that the assumption of randomly placed nodes is crucial for this conclusion to be valid, as was discussed in detail in Section 3.

4.8 Chapter Summary

In this chapter, we have obtained information-theoretic inner and outer bounds on the n^2-dimensional unicast capacity region of a wireless network with n randomly placed nodes and assuming a Gaussian fading channel model. These bounds are tight (in the scaling sense) along $n^2 - n$ of the total n^2 dimensions for $\alpha \in (2, 5]$ (corresponding to balanced traffic), and along all n^2 dimensions for $\alpha > 5$. This approximate characterization is in terms of $2n$ weighted cuts, which are based on the geometry of the locations of the source nodes and their destination nodes and on the traffic demands between them, and thus can be readily evaluated.

This characterization is obtained by establishing that, for balanced traffic or for $\alpha > 5$, the unicast capacity region of a capacitated tree graph under routing has the

same scaling as the unicast capacity region of the original wireless network. The leaf nodes of this tree graph correspond to the nodes in the wireless network, and internal nodes of the tree graph correspond to hierarchically growing sets of nodes.

This equivalence suggests a three-layer communication architecture for achieving the entire unicast capacity region (in the scaling sense). The top or routing layer establishes paths from each of the source nodes to its destination over the tree graph. The middle or cooperation layer provides this tree abstraction to the routing layer by distributing the traffic among the corresponding subset of nodes as a message travels up the tree graph, and by concentrating the traffic on to the corresponding subset of nodes as the message travels down the tree. The bottom or physical layer implements this distribution and concentration of traffic over the wireless network. The implementation of this distribution and concentration of traffic depends on the path-loss exponent: For low path-loss exponent ($\alpha \in (2, 3]$), hierarchical relaying is used, while for high path-loss exponent ($\alpha > 3$), multi-hop communication is used.

Chapter 5

Service Heterogeneity: Multicast

In this chapter, we analyze the scaling of the $n \times 2^n$-dimensional multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ of a wireless network of n randomly placed nodes under a Gaussian fading channel model.

We first present an inner and an outer bound to the $n \times 2^n$-dimensional multicast capacity region. We show that the two bounds coincide (up to scaling) along $n2^n - n$ dimensions (corresponding to balanced traffic) for $\alpha \in (2, 5]$ and along all $n2^n$ dimensions for $\alpha > 5$. We show that, as the unicast capacity region, the multicast capacity region can be approximated by a polytope with less than $2n$ faces, each corresponding to a distinct cut (i.e., a subset of nodes) in the wireless network. Again, as in the unicast case, only $2n$ out of 2^n possible cuts in the wireless network are asymptotically relevant.

Second, we show how the three-layer communication architecture introduced in Chapter 4 can be adapted for the transmission of multicast traffic. Recall the three layers of this scheme: The routing layer, the cooperation layer, and the physical layer. We show that only the routing layer needs to be changed to accommodate multicast traffic. The other two layers are unaffected. The approximate optimality of this three-layer architecture implies that a separation based approach, in which routing is performed independently of the physical layer, is order-optimal for balanced multicast traffic or for $\alpha > 5$, and hence techniques such as network coding are not necessary for order-optimal transmission of multicast traffic in wireless networks.

5.0.1 Organization

The remainder of this chapter is organized as follows. Section 5.1 presents the main results of this chapter. In Section 5.2, we analyze various examples scenarios with heterogeneous multicast traffic patterns for which no scaling results were previously known. Section 5.3 provides a high-level description of the modifications to the three-layer architecture required for multicast traffic. Section 5.4 contains the proofs of the main results. Finally, Sections 5.5 and 5.6 contain discussions and concluding remarks.

5.1 Main Results

This section contains the main results of this chapter. In Section 5.1.1, we provide inner and outer bounds on the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ and conditions for these bounds to be tight in the scaling sense. In Section 5.1.2, we discuss implications of these results on the behavior of the multicast capacity region for large values of n. We consider computational aspects in Section 5.1.3.

5.1.1 Multicast Capacity Region

Recall the definition of $\widetilde{L}(n)$ in (4.1) and of the subsets $\{V_{\ell,i}(n)\}_{i=1}^{4^\ell}$ of $V(n)$ at level ℓ in Section 4.1.1. Define

$$\begin{aligned}\widehat{\Lambda}_1^{\mathsf{MC}}(n) \triangleq \Big\{\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n\times 2^n} :\ & \sum_{u\in V_{\ell,i}(n)} \sum_{\substack{W\subset V(n):\\ W\cap V_{\ell,i}^c(n)\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{u\in V_{\ell,i}^c(n)} \sum_{\substack{W\subset V(n):\\ W\cap V_{\ell,i}(n)\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} \leq (4^{-\ell}n)^{2-\min\{3,\alpha\}/2} \\ & \qquad \forall \ell \in \{1,\ldots,\widetilde{L}(n)\}, i\in\{1,\ldots 4^\ell\}, \\ & \sum_{\substack{W\subset V(n):\\ W\setminus\{u\}\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{\tilde{u}\neq u} \sum_{\substack{W\subset V(n):\\ u\in W}} \lambda_{\tilde{u},W}^{\mathsf{MC}} \leq 1 \\ & \qquad \forall u \in V(n)\Big\},\end{aligned} \tag{5.1}$$

and

$$\widehat{\Lambda}_2^{\mathsf{MC}}(n) \triangleq \Big\{\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n\times 2^n} : \sum_{u\in V_{\ell,i}(n)} \sum_{\substack{W\subset V(n):\\ W\cap V_{\ell,i}^c(n)\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} \leq (4^{-\ell}n)^{2-\min\{3,\alpha\}/2} \quad \forall \ell \in \{1,\ldots,\widetilde{L}(n)\}, i \in \{1,\ldots 4^\ell\}, \sum_{\substack{W\subset V(n):\\ W\setminus\{u\}\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{\tilde{u}\neq u} \sum_{\substack{W\subset V(n):\\ u\in W}} \lambda_{\tilde{u},W}^{\mathsf{MC}} \leq 1 \quad \forall u \in V(n)\Big\}.$$

The definitions of $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$ are similar to those of $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$ in Chapter 4. $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$ are the collection of all multicast traffic matrices λ^{MC} such that for various cuts $S \subset V(n)$ in the network, the total traffic demand (in either one or both directions)

$$\sum_{u\in S} \sum_{\substack{W\subset V(n):\\ W\cap S^c\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}}$$

$$\sum_{u\in S} \sum_{\substack{W\subset V(n):\\ W\cap S^c\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{u\in S^c} \sum_{\substack{W\subset V(n):\\ W\cap S\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}}$$

across the cut S is not too big. Note that, unlike in the definitions of $\widehat{\Lambda}_1^{\mathsf{UC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{UC}}(n)$, we count $\lambda_{u,W}$ as crossing the cut S (in the outgoing direction, say) if $u \in S$ and $W \cap S^c \neq \emptyset$, i.e., if there is at least one node w in the multicast destination group W that lies outside S. The number of such cuts S we need to consider is at most $2n$, as in the unicast case.

The next theorem shows that $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ is an approximate (in the scaling sense) inner bound and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$ is an approximate outer bound to the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ of the wireless network.

Theorem 5.1. *Under either fast or slow fading, for any $\alpha > 2$, there exist*

$$b_1(n) \geq n^{-o(1)},$$
$$b_2(n) = O(\log^6(n)),$$

such that

$$b_1(n)\widehat{\Lambda}_1^{\mathsf{MC}}(n) \subset \Lambda^{\mathsf{MC}}(n) \subset b_2(n)\widehat{\Lambda}_2^{\mathsf{MC}}(n),$$

with probability $1 - o(1)$ *as* $n \to \infty$.

Theorem 5.1 holds only with probability $1 - o(1)$ for different reasons for the fast and slow fading cases. Under fast fading, the theorem holds for almost all node placements. Under slow fading, the theorem holds under the same conditions on the node placement, but now it also only holds for almost all realization of the fading $\{\theta_{u,v}\}_{u,v}$.

Comparing the expressions for $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$, we see that whenever a traffic matrix λ^{MC} satisfies

$$\sum_{\substack{W \subset V(n): \\ W \cap V_{\ell,i}^c(n) \neq \emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{u \in V_{\ell,i}^c(n)} \sum_{\substack{W \subset V(n): \\ W \cap V_{\ell,i}(n) \neq \emptyset}} \lambda_{u,W}^{\mathsf{MC}} \leq n^{o(1)} \sum_{\substack{W \subset V(n): \\ W \cap V_{\ell,i}^c(n) \neq \emptyset}} \lambda_{u,W}^{\mathsf{MC}} \tag{5.2}$$

for all $\ell \in \{1, \ldots, \widetilde{L}(n)\}$, $i \in \{1, \ldots 4^\ell\}$ then $\lambda^{\mathsf{MC}} \in \widehat{\Lambda}_2^{\mathsf{MC}}(n)$ implies $n^{-o(1)}\lambda^{\mathsf{MC}} \in \widehat{\Lambda}_1^{\mathsf{MC}}(n)$, and hence, for such traffic matrices, the inner and outer bound in Theorem 5.1 coincide up to scaling. In particular, this applies for multicast traffic matrices λ^{MC} such that (5.2) holds with equality. We call such traffic *(approximately) balanced* in the following. Note that the the condition of balanced multicast traffic imposes (at most) n linear constraints on λ^{MC}, and hence the inner and outer bound in Theorem 5.1 are tight up to scaling along at least $n2^n - n$ out of all $n2^n$ total dimensions of $\Lambda^{\mathsf{MC}}(n)$.

For high path-loss exponent ($\alpha > 5$), the next theorem shows that the inner bound $\Lambda_1^{\mathsf{MC}}(n)$ in Theorem 5.1 is also an approximate outer bound to $\Lambda^{\mathsf{MC}}(n)$.

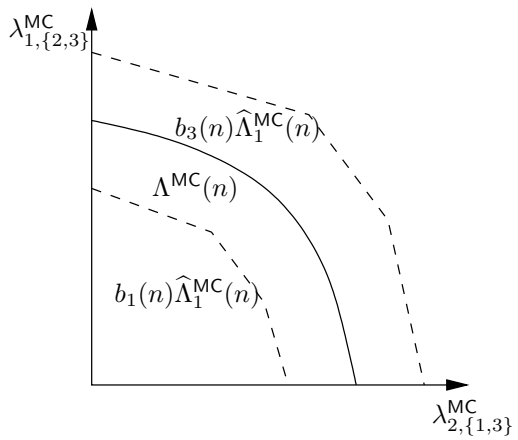

Figure 5-1: For $\alpha > 5$, the set $\widehat{\Lambda}^{\mathsf{MC}}_1(n)$ approximates the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ of the wireless network in the sense that $b_1(n)\widehat{\Lambda}^{\mathsf{MC}}_1(n)$ (with $b_1(n) \geq n^{-o(1)}$) provides an inner bound to $\Lambda^{\mathsf{MC}}(n)$ and $b_3(n)\widehat{\Lambda}^{\mathsf{MC}}_1(n)$ (with $b_3(n) = O\big(\log^6(n)\big)$) provides an outer bound to $\Lambda^{\mathsf{MC}}(n)$. The figure shows two dimensions (namely $\lambda^{\mathsf{MC}}_{1,\{2,3\}}$ and $\lambda^{\mathsf{MC}}_{2,\{1,3\}}$) of the $n \times 2^n$-dimensional set $\Lambda^{\mathsf{MC}}(n)$. The same approximation result holds for $\alpha \in (2, 5]$ along at least $n2^n - n$ out of $n2^n$ dimensions.

Theorem 5.2. *Under either fast or slow fading, for any $\alpha > 5$, there exists*

$$b_3(n) = O(\log^6(n)),$$

such that

$$\Lambda^{\mathsf{MC}}(n) \subset b_3(n)\widehat{\Lambda}^{\mathsf{MC}}_1(n),$$

with probability $1 - o(1)$ as $n \to \infty$.

Theorems 5.1 and 5.2 imply that the quantity $\widehat{\Lambda}^{\mathsf{MC}}_1(n)$ determines the scaling of the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ for $\alpha > 5$ and along all dimensions corresponding to balanced traffic for $\alpha \in (2, 5]$. This is illustrated in Figure 5-1. The approximation is within a factor $n^{\pm o(1)}$, which, as in the unicast case, can be further sharpened (see Section 4.7.2 in Chapter 4 for a discussion).

5.1.2 Implications of Theorems 5.1 and 5.2

As in the unicast case, Theorems 5.1 and 5.2 can be applied in two ways. First, the theorem can be used to analyze the asymptotic achievability of a sequence of multicast traffic matrices. Let $\{\lambda^{\mathsf{MC}}(n)\}_{n\geq 1}$ be such a sequence of multicast traffic matrices with $\lambda^{\mathsf{MC}}(n) \in \mathbb{R}_+^{n\times 2^n}$. Define

$$\rho^*_{\lambda^{\mathsf{MC}}}(n) \triangleq \sup\{\rho : \rho\lambda^{\mathsf{MC}}(n) \in \Lambda^{\mathsf{MC}}(n)\},$$
$$\hat{\rho}^*_{\lambda^{\mathsf{MC}}}(n) \triangleq \sup\{\hat{\rho} : \hat{\rho}\lambda^{\mathsf{MC}}(n) \in \widehat{\Lambda}^{\mathsf{MC}}(n)\},$$

in analogy to the unicast case. Theorems 5.1 and 5.2 state that if either $\alpha > 5$ or all $\lambda^{\mathsf{MC}}(n)$ are balanced then[1]

$$\lim_{n\to\infty} \frac{\log(\rho^*_{\lambda^{\mathsf{MC}}}(n))}{\log(n)} = \lim_{n\to\infty} \frac{\log(\hat{\rho}^*_{\lambda^{\mathsf{MC}}}(n))}{\log(n)}.$$

Several applications of this approach are explored in Section 5.2.

Second, Theorems 5.1 and 5.2 provide information about the shape of the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$. As in the unicast case, the boundary points of $\Lambda^{\mathsf{MC}}(n)$ vary at least from $n^{-\min\{3,\alpha\}/2+o(1)}$ to $n^{-o(1)}$, and this variation on exponential scale is preserved by $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ and $\widehat{\Lambda}_2^{\mathsf{MC}}(n)$. See also Section 4.1.2 for more details.

5.1.3 Computational Aspects

In this section, we show that $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ can be efficiently described. By Theorems 5.1 and 5.2 this yields a computationally efficient approximate description of the entire multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ for $\alpha > 5$ and of $n2^n - n$ of its $n2^n$ dimensions for $\alpha \in (2, 5]$.

The multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ is a $n \times 2^n$-dimensional set, i.e., the number of dimensions is exponentially large in n. Nevertheless, its approximation $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$ can (as in the unicast case) be computed by evaluating at most $2n$ cuts. This yields

[1] We again assume that the limits exist, otherwise the same statement holds for lim sup and lim inf.

a very compact approximate representation of the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ (i.e., we represent a region of exponential size in n as an intersection of only linearly many halfspaces — one halfspace corresponding to each cut). Moreover, it implies that membership $\lambda^{\mathsf{MC}} \in \widehat{\Lambda}_1^{\mathsf{MC}}(n)$ can be computed efficiently. More precisely, evaluating each of the $\Theta(n)$ cuts takes at most $|\{(u,W) : \lambda_{u,W}^{\mathsf{MC}} > 0\}|$ operations. Thus membership $\lambda^{\mathsf{MC}} \in \widehat{\Lambda}_1^{\mathsf{MC}}(n)$ can be tested in at most $\Theta(n)$ times more operations than required to just read the problem parameters. In other words, we have a linear time (in the length of the input) algorithm for testing membership of a multicast traffic matrix λ^{MC} in $\widehat{\Lambda}_1^{\mathsf{MC}}(n)$, and hence for $\alpha > 5$ or balanced λ^{MC} also for approximate testing of membership in $\Lambda^{\mathsf{MC}}(n)$. However, this algorithm is not necessarily linear time in n since reading just the input $\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n\times 2^n}$ itself might take exponential time in n.

5.2 Example Scenarios

Here we consider several multicast scenarios. The first two examples consider broadcast traffic. The third and fourth examples consider proper multicast traffic (i.e., only a subset of nodes acts as destinations).

Example 5.1. *Broadcast from one source*

Assume we have only one source (say $u_0 \in V(n)$) that wants to broadcast the same message to all other nodes. In other words, we consider the multicast traffic matrix

$$\lambda_{u,W}^{\mathsf{MC}} = \begin{cases} \rho(n) & \text{if } u = u_0 \text{ and } W = V(n), \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) > 0$. Applying Theorem 5.1 yields that $\rho^*(n)$, the largest achievable $\rho(n)$, satisfies

$$\rho^*(n) = n^{\pm o(1)}$$

with probability $1 - o(1)$ as $n \to \infty$. This implies that the source can broadcast its information at essentially constant rate independent of n to all nodes in the network.

Hence, broadcasting information from one source to all nodes in the network is (at least asymptotically) no harder than transmitting information from one source to its destination. ◊

Example 5.2. *Broadcast from many sources*

Consider a scenario with n^β sources, $\{u_1, \ldots, u_{n^\beta})$, for some $0 \leq \beta \leq 1$, each broadcasting an independent message to all other nodes at the same rate. In other words, we have a multicast traffic matrix of the form

$$\lambda_{u,W}^{\mathsf{MC}} = \begin{cases} \rho(n) & \text{if } u = u_i \text{ for some } i \text{ and } W = V(n), \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) > 0$. Applying Theorem 5.1 yields that $\rho^*(n)$, the largest achievable $\rho(n)$, satisfies

$$\rho^*(n) = n^{-\beta \pm o(1)},$$

with probability $1 - o(1)$ as $n \to \infty$. ◊

Example 5.3. *Multicast from many sources*

Assume each node generates independent multicast traffic for a set of n^β destinations. All these sources and their destinations are chosen independently and uniformly at random from $V(n)$. Let $\{u_1, \ldots, u_n\}$ be the source nodes and $\{W_1, \ldots, W_n$ with $|W_i| = n^\beta$ the corresponding multicast nodes. Then the multicast traffic matrix λ^{MC} is of the form

$$\lambda_{u,W}^{\mathsf{MC}} = \begin{cases} \rho(n) & \text{if } u = u_i, W = W_i \text{ for some } i, \\ 0 & \text{else,} \end{cases} \tag{5.3}$$

for some $\rho(n) > 0$. This traffic matrix is approximately balanced with high probability, and hence applying Theorem 5.1 yields that $\rho^*(n)$ satisfies

$$\rho^*(n) = \min\left\{n^{\pm o(1)}, n^{(1-\beta)(2-\bar{\alpha}/2)-1 \pm o(1)}\right\} \tag{5.4}$$

with probability $1 - o(1)$ as $n \to \infty$, and with

$$\bar{\alpha} \triangleq \min\{3, \alpha\}.$$

Note that for $\beta_2 = 0$, we recover the result for unicast traffic with random source-destination pairing and uniform rate. $\Diamond$

Example 5.4. *Localized multicast from many sources*

Consider the setup of Example 5.3 above, except now each source picks n^{β_1} destinations uniformly at random from among nodes within a distance of $n^{\frac{\beta_2}{2}}$, where $\beta_2 > \beta_1$. In other words, each source node performs localized multicast. Again, let $\{u_1, \ldots, u_n\}$ denote the source nodes and $\{W_1, \ldots, W_n\}$, with $|W_i| = n^{\beta_1}$, denote the corresponding destination nodes, where now $r_{u_i,v} \leq n^{\frac{\beta_2}{2}}$, for each $v \in W_i$. The multicast traffic matrix λ^{MC} is of the form (5.3). This traffic matrix is again approximately balanced with high probability, and an application of Theorem 5.1 shows that $\rho^*(n)$ satisfies

$$\rho^*(n) = \min\left\{n^{\pm o(1)}, n^{(\beta_2-\beta_1)(2-\bar{\alpha}/2)-\beta_2\}\pm o(1)}\right\}$$

with probability $1 - o(1)$ as $n \to \infty$. Note that setting $\beta_2 = 1$, i.e., each source picks its destinations uniformly over the entire region $A(n)$, yields the same scaling of $\rho^*(n)$ as in Example 5.3. Similarly, for $\beta_1 = 0$, and $\beta_2 = \beta$, we recover the unicast scenario with $K = 1$ in Example 4.1. $\Diamond$

5.3 Communication Scheme for Multicast Traffic

Here we show that the same communication scheme presented in Section 4.3 for general unicast traffic can also be used to transmit general multicast traffic. It is the tree structure of the scheme that is critically exploited in the proof of Theorem 5.1 to obtain an approximation of the multicast capacity region $\Lambda^{\mathsf{MC}}(n)$.

We will use the same three-layer architecture as for unicast traffic presented in Section 4.3. Note that for this, we only need to modify the operation of the top or

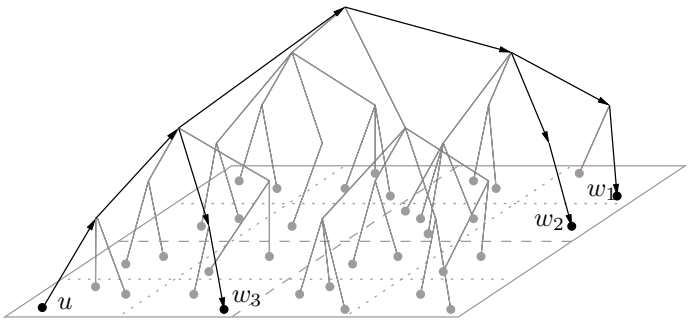

Figure 5-2: Example operation of the routing layer in the three-layer architecture under multicast traffic.

routing layer. Indeed, no matter how we configure the routing layer, the lower layers operate as before.

We now outline how the routing layer needs to be adapted for the multicast case. Consider a multicast message that needs to be transmitted from a source node $u \in V(n)$ to its set of intended destinations $W \subset V(n)$. In the routing layer, we want to route this message from u to W over G. Since G is a tree, the routing part is simple. In fact, between u and every $w \in W$ there exists a unique path in G. Consider the union of all those paths. It is easy to see that this union is a subtree of G. Indeed, it is the smallest subtree of G that covers $\{u\} \cup W$. Traffic is optimally routed over G from u to W by sending it along the edges of this subtree.

The next example illustrates the operation of the routing layer under multicast traffic.

Example 5.5. Consider one source node u and the corresponding multicast group $W \triangleq \{w_1, w_2, w_3\}$ as shown in Figure 5-2.

In the routing layer, we find the smallest subgraph $G(\{u\} \cup W)$ covering $\{u\} \cup W$ (indicated by black lines in Figure 5-2). Messages are sent from the source to its destinations by routing them along this subgraph. In other words, $G(\{u\} \cup W)$ is the multicast tree along which the message is sent from u to W. The cooperation layer and physical layer operate in the same way as for unicast traffic (see Figure 4-4 for an example). $\Diamond$

5.4 Proofs

This Section contains the proofs of Theorem 5.1 (in Section 5.4.1) and of Theorem 5.2 (in Section 5.4.2).

5.4.1 Proof of Theorem 5.1

The proof of Theorem 5.1 relies on linking the multicast capacity region to the unicast capacity region. We say that a unicast traffic matrix λ^{UC} is *compatible* with a multicast traffic matrix λ^{MC} if there exists a mapping $f : V(n) \times 2^{V(n)} \to V(n)$ such that $f(u, W) \in W$, for all (u, W), and

$$\lambda^{\mathsf{UC}}_{u,v} = \sum_{\substack{W \subset V(n): \\ f(u,W)=v}} \lambda^{\mathsf{MC}}_{u,W}$$

for all (u, v). In words, λ^{MC} is compatible with λ^{UC} if we can create the unicast traffic matrix λ^{UC} from λ^{MC} by simply discarding the traffic for the pair (u, W) at all the nodes $W \setminus \{f(u, W)\}$. Let $\Gamma(\lambda^{\mathsf{MC}})$ be the set of all unicast traffic matrices compatible with the multicast traffic matrix λ^{MC}. We extend the definition of Γ to sets of traffic matrices as follows:

$$\Gamma\big(\Lambda^{\mathsf{MC}}(n)\big) \triangleq \bigcup_{\lambda^{\mathsf{MC}} \in \Lambda^{\mathsf{MC}}(n)} \Gamma(\lambda^{\mathsf{MC}}),$$
$$\Gamma^{-1}\big(\Lambda^{\mathsf{UC}}(n)\big) \triangleq \big\{\lambda^{\mathsf{MC}} : \Gamma(\lambda^{\mathsf{MC}}) \subset \Lambda^{\mathsf{UC}}(n)\big\}.$$

In words, $\Gamma\big(\Lambda^{\mathsf{MC}}(n)\big)$ is the collection of all unicast traffic matrices that are compatible with a multicast traffic in $\Lambda^{\mathsf{MC}}(n)$, and $\Gamma^{-1}\big(\Lambda^{\mathsf{UC}}(n)\big)$ is the collection of all multicast traffic matrices that have all their compatible unicast traffic matrices in $\Lambda^{\mathsf{UC}}(n)$.

We start with some auxiliary lemmas for the outer bound in Theorem 5.1.

Lemma 5.3. *Under either fast or slow fading, for any $\alpha > 2$,*

$$\Lambda^{\mathsf{MC}}(n) \subset \Gamma^{-1}\big(\Lambda^{\mathsf{UC}}(n)\big).$$

Proof. It is clear that if $\lambda^{\mathsf{MC}} \in \Lambda^{\mathsf{MC}}$ then $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ for any $\lambda^{\mathsf{UC}} \in \Gamma(\lambda^{\mathsf{MC}})$. Indeed, we can reliably transmit at rate λ^{UC} by using the communication scheme for λ^{MC} and simply discarding the messages delivered by this scheme at all but one node in each multicast destination group. Thus $\lambda^{\mathsf{MC}} \in \Lambda^{\mathsf{MC}}$ implies $\Gamma(\lambda^{\mathsf{MC}}) \subset \Lambda^{\mathsf{UC}}$, and therefore

$$\Gamma\big(\Lambda^{\mathsf{MC}}\big) = \bigcup_{\lambda^{\mathsf{MC}} \in \Lambda^{\mathsf{MC}}} \Gamma\big(\lambda^{\mathsf{MC}}\big) \subset \Lambda^{\mathsf{UC}},$$

or, equivalently,

$$\Lambda^{\mathsf{MC}} \subset \Gamma^{-1}\big(\Lambda^{\mathsf{UC}}\big),$$

proving the lemma. □

We now prove several auxiliary results for the inner bound in Theorem 5.1. Consider again the tree graph $G = (V_G, E_G)$ with leaf nodes $V(n) \subset V_G$ constructed in Section 4.5. As before, we consider traffic between leaf nodes of G. In particular, any multicast traffic matrix $\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n \times 2^n}$ for the wireless network is also a multicast traffic matrix for the graph G. Denote by $\Lambda_G^{\mathsf{MC}}(n) \subset \mathbb{R}_+^{n \times 2^n}$ the set of feasible (under routing) multicast traffic matrices between leaf nodes of G.

The next lemma shows that if multicast traffic can be routed over G then approximately the same multicast traffic can be transmitted reliably over the wireless network. Before we state that lemma, recall the definition of the set $\mathcal{V}(n)$ of node placements $V(n)$ that satisfy certain regularity conditions as defined in Section 4.4.1.

Lemma 5.4. *Under fast fading, for any $\alpha > 2$, there exists $b(n) \geq n^{-o(1)}$ such that for all $V(n) \in \mathcal{V}(n)$,*

$$b(n)\Lambda_G^{\mathsf{MC}}(n) \subset \Lambda^{\mathsf{MC}}(n).$$

The same statement holds under slow fading with probability $1 - o(1)$ as $n \to \infty$.

Proof. The proof follows using the same construction as in Lemma 4.10. □

We now show that since G is a tree graph, the multicast capacity region $\Lambda_G^{\mathsf{MC}}(n)$ can be described through the unicast capacity region $\Lambda_G^{\mathsf{UC}}(n)$ of G. The fact that G

is a tree is critical for this result to hold.

Lemma 5.5. *For any $\alpha > 2$,*

$$\Gamma^{-1}\big(\Lambda_G^{\mathsf{UC}}(n)\big) \subset \Lambda_G^{\mathsf{MC}}(n).$$

Proof. Assume that $\lambda^{\mathsf{MC}} \notin \Lambda_G^{\mathsf{MC}}$. Since G is a tree, there is only one way to route multicast traffic from u to W, namely along the subtree $G(\{u\} \cup W)$ induced by $\{u\} \cup W$ (i.e., the smallest subtree of G that covers $\{u\} \cup W$). Hence for any edge $e \in E_G$, the traffic $d_{\lambda^{\mathsf{MC}}}(e)$ that needs to be routed over e is equal to

$$d_{\lambda^{\mathsf{MC}}}(e) = \sum_{\substack{u \in V, W \subset V: \\ e \in E_{G(\{u\} \cup W)}}} \lambda_{u,W}^{\mathsf{MC}}.$$

Now, since $\lambda^{\mathsf{MC}} \notin \Lambda_G^{\mathsf{MC}}$, there exists $e \in E_G$ such that

$$d_{\lambda^{\mathsf{MC}}}(e) > c_e.$$

But then, by definition of $d_{\lambda^{\mathsf{MC}}}(e)$, there exists a function $f : V \times 2^V \to V$ with $f(u, W) \in W$ for all (u, W), and such that

$$d_{\lambda_f^{\mathsf{UC}}(\lambda^{\mathsf{MC}})}(e) = d_{\lambda^{\mathsf{MC}}}(e) > c_e,$$

where $\lambda_f^{\mathsf{UC}}(\lambda^{\mathsf{MC}})$ is the unicast traffic matrix resulting from applying f to the multicast traffic matrix λ^{MC}. Thus $\lambda_f^{\mathsf{UC}}(\lambda^{\mathsf{MC}}) \notin \Lambda_G^{\mathsf{UC}}$.

We have shown that if $\lambda^{\mathsf{MC}} \notin \Lambda_G^{\mathsf{MC}}$ then there exists $\lambda^{\mathsf{UC}} \in \Gamma(\lambda^{\mathsf{MC}})$ such that $\lambda^{\mathsf{UC}} \notin \Lambda_G^{\mathsf{UC}}$. In other words, $\Gamma(\lambda^{\mathsf{MC}})$ is not a subset of Λ_G^{UC}, and therefore $\lambda^{\mathsf{MC}} \notin \Gamma^{-1}\big(\Lambda_G^{\mathsf{UC}}\big)$. Hence $\lambda^{\mathsf{MC}} \notin \Lambda_G^{\mathsf{MC}}$ implies $\lambda^{\mathsf{MC}} \notin \Gamma^{-1}\big(\Lambda_G^{\mathsf{UC}}\big)$, and thus

$$\Gamma^{-1}\big(\Lambda_G^{\mathsf{UC}}\big) \subset \Lambda_G^{\mathsf{MC}}.$$

□

We are now ready for the proof of Theorem 5.1. Note that

$$\Gamma(\lambda^{\mathsf{MC}}) \subset \widehat{\Lambda}_j^{\mathsf{UC}}(n) \quad \Leftrightarrow \quad \lambda^{\mathsf{MC}} \in \widehat{\Lambda}_j^{\mathsf{MC}}(n),$$

and hence

$$\Gamma^{-1}\Big(\widehat{\Lambda}_j^{\mathsf{UC}}(n)\Big) = \widehat{\Lambda}_j^{\mathsf{MC}}(n), \tag{5.5}$$

for $j \in \{1, 2\}$.

For the inner bound in Theorem 4.1, we have for $V \in \mathcal{V}$

$$\begin{aligned} b_1(n)\widehat{\Lambda}_1^{\mathsf{MC}}(n) &= b_1(n)\Gamma^{-1}\Big(\widehat{\Lambda}_1^{\mathsf{UC}}(n)\Big) && \text{(5.6a)} \\ &= b_1(n)\Gamma^{-1}\big(\Lambda_G^{\mathsf{UC}}(n)\big) && \text{(5.6b)} \\ &\subset b_1(n)\Lambda_G^{\mathsf{MC}}(n) && \text{(5.6c)} \\ &\subset \Lambda^{\mathsf{MC}}(n) && \text{(5.6d)} \end{aligned}$$

for $b_1(n) \geq n^{-o(1)}$, and where (5.6a) follows from (5.5), (5.6b) follows from Lemma 4.11, (5.6c) follows from Lemma 5.5, and (5.6d) follows from Lemma 5.4.

For the outer bound in Theorem 4.1, we have for $V \in \mathcal{V}$

$$\begin{aligned} \Lambda^{\mathsf{MC}}(n) &\subset \Gamma^{-1}\big(\Lambda^{\mathsf{UC}}(n)\big) && \text{(5.7a)} \\ &\subset b_2(n)\Gamma^{-1}\Big(\widehat{\Lambda}_2^{\mathsf{UC}}(n)\Big) && \text{(5.7b)} \\ &= b_2(n)\widehat{\Lambda}_2^{\mathsf{MC}}(n) && \text{(5.7c)} \end{aligned}$$

for $b_2(n) = O(\log^6(n))$, and where (5.7a) follows from Lemma 5.3, (5.7b) follows from Theorem 4.1, and (5.7c) follows from (5.5).

Since $V \in \mathcal{V}$ with probability $1 - o(1)$ by Lemma 4.5, this proves Theorem 5.1.

5.4.2 Proof of Theorem 5.2

For $V \in \mathcal{V}$, and $\alpha > 5$,

$$\Lambda^{\mathsf{MC}}(n) \subset \Gamma^{-1}\big(\Lambda^{\mathsf{UC}}(n)\big) \tag{5.8a}$$

$$\subset b_3(n)\Gamma^{-1}\Big(\widehat{\Lambda}_1^{\mathsf{UC}}(n)\Big) \tag{5.8b}$$

$$= b_3(n)\widehat{\Lambda}_1^{\mathsf{MC}}(n) \tag{5.8c}$$

for $b_2(n) = O(\log^6(n))$, and where (5.8a) follows from Lemma 5.3, (5.8b) follows from Theorem 4.2, and (5.8c) follows from (5.5). Since $V \in \mathcal{V}$ with probability $1 - o(1)$ by Lemma 4.5, this proves Theorem 5.2.

5.5 Discussion

We discuss several aspects and extensions of the modified three-layer architecture introduced in Section 5.3 for transmission of multicast traffic. In Section 5.5.1, we show how the results for multicast traffic in extended networks can be used to obtain scaling results for multicast traffic in dense networks. In Section 5.5.2, we discuss design guidelines for the transmission of multicast traffic in large wireless networks.

5.5.1 Dense Networks

Up to this point, we have only considered multicast traffic in *extended* networks, i.e., n nodes are located on a square of area n. As in previous chapters, we now briefly sketch how these results can be recast for *dense* networks, in which n nodes are located on a square of unit area. As in the unicast case, the scaling of the multicast capacity region for dense networks can be obtained from the one for extended networks by taking a limit as $\alpha \to 2$.

The resulting approximate multicast capacity regions $\widehat{\Lambda}^{\mathsf{MC}}(n)$ has again a particularly simple shape in this limit. As in the unicast case, the only constraints in (5.1) that can be tight are at level $\ell = \log(n)$. This results in the following approximate

multicast capacity region for dense networks:

$$\widehat{\Lambda}_1^{\mathsf{MC}}(n) \triangleq \Big\{\lambda^{\mathsf{MC}} \in \mathbb{R}_+^{n\times 2^n} : \sum_{\substack{W\subset V(n):\\ W\setminus\{u\}\neq\emptyset}} \lambda_{u,W}^{\mathsf{MC}} + \sum_{\tilde{u}\neq u} \sum_{\substack{W\subset V(n):\\ u\in W}} \lambda_{\tilde{u},W}^{\mathsf{MC}} \leq 1,\ \forall u \in V(n)\Big\},$$

and we obtain that for dense networks, for any $\alpha > 2$,

$$n^{-o(1)}\widehat{\Lambda}_1^{\mathsf{MC}}(n) \subset \Lambda^{\mathsf{MC}}(n) \subset O(\log^6(n))\widehat{\Lambda}_1^{\mathsf{MC}}(n).$$

Note that, in contrast to the extended case, this results in an approximate characterization of the entire unicast region for all $\alpha > 2$.

5.5.2 Design Guidelines

Several design guidelines for multicast traffic in large wireless networks can be obtained from the results presented in this chapter. First, observe that the only modification to the three-layer architecture introduced in Chapter 4 for unicast traffic is in the routing layer. This suggests that multicasting in wireless networks can be handled by appropriate routing alone. In particular, for balanced multicast traffic or $\alpha > 5$, cross-layer techniques such as network coding (which can provide an arbitrarily large increase in achievable rates in wireline networks) are not necessary for order-optimal transmission of multicast traffic in wireless networks, and can provide at most a factor $n^{o(1)}$ increase in achievable rates.

Second, note that the problem of finding multicast trees in wireless networks is solved very efficiently in our proposed three-layer architecture. Namely, we map the wireless network to the graph G and then look for multicast trees on G. Since G is a tree, finding the optimal multicast trees in G is trivial. The optimality of the three-layer architecture in the scaling sense for balanced traffic or $\alpha > 5$ suggests this approach for multicast routing in large wireless networks.

Finally, we point out that this approach of finding optimal multicast trees in G is also useful in streaming applications, in which the nodes that are subscribed to a certain multicast stream can vary over time. In fact, the optimal multicast tree

connecting source u with multicast group W in G is $G(\{u\} \cup W)$ (i.e., the smallest subtree of G connecting all nodes in $\{u\} \cup W$). Now, note that $G(\{u\} \cup W)$ has the property that if $W = \{w_1, \ldots, w_m\}$ then

$$G(\{u\} \cup W) = \bigcup_{i=1}^{m} G(\{u, w_i\}),$$

i.e., the optimal multicast tree connecting the source u with its subscribers W can be constructed as the union of all the paths between u and w_i. This implies that if a new node subscribes to a stream available at u, the optimal multicast tree can be updated efficiently by just adding this new path (and similar if a node unsubscribes from a stream).

5.6 Chapter Summary

In this chapter, we have obtained inner and outer bounds for the $n \times 2^n$-dimensional multicast capacity region of a wireless network with n randomly placed nodes and assuming a Gaussian fading channel model. These inner and outer bounds coincide (up to scaling) along at least $n2^n - n$ out of $n2^n$ dimensions for $\alpha \in (2, 5]$ and in all $n2^n$ dimensions for $\alpha > 5$. As in the unicast case, this approximate characterization of the multicast capacity region is in terms of $2n$ weighted cuts. This provides a compact approximate representation of the multicast capacity region.

We have shown how the three-layer communication architecture introduced for general unicast traffic can be modified for multicast traffic. Out of the three layers, only the top or routing layer needed to be changed — the other layers operate as in the unicast case.

This scheme also establishes that a separation based approach, where the routing layer works essentially independently of the physical layer, is order optimal for balanced traffic or when $\alpha > 5$. Thus, such techniques as network coding can provide at most a small increase in the scaling.

Chapter 6

Service Heterogeneity: Caching

In this chapter, we analyze the scaling of the $2^n \times n$-dimensional caching capacity region $\Lambda^{\mathsf{CA}}(n)$ under random node placement. We present an achievable communication scheme for the caching problem, yielding an inner bound on the caching capacity region. For large values of path-loss exponent, we provide a matching (in the scaling sense) outer bound, proving the optimality (again in the scaling sense) of our proposed scheme. Together, this provides a scaling description of the entire caching capacity region of the wireless network in the large path-loss regime. The proposed communication scheme solves the problem of optimal cache selection and channel coding separately, showing that such a separation is order-optimal.

6.0.1 Organization

The remainder of the chapter is organized as follows. In Section 6.1, we present the main results of this chapter. We analyze several example scenarios in Section 6.2. In Section 6.3, we introduce the communication scheme achieving the inner bound on $\Lambda^{\mathsf{CA}}(n)$. Section 6.4 contains proofs. Sections 6.5 and 6.6 contain discussions and concluding remarks.

6.1 Main Results

In Section 6.1.1, we provide an inner and a matching (in the scaling sense) outer bound on the capacity region $\Lambda^{\mathsf{CA}}(n)$. In Section 6.1.2, we discuss computational aspects.

6.1.1 Caching Capacity Region

Recall the construction of the graph $G = (V_G, E_G)$ introduced in Section 4.3. G is a tree with leaf nodes $V(n) \subset V_G$. Leaf nodes in G share the same parent node in G if they fall within the same grid square at level $\widetilde{L}(n)$ in $A(n)$ (with $\widetilde{L}(n)$ as defined in (4.1)). Nodes at level ℓ in the tree G share the same parent node if all their children fall in the same grid square at level $\ell - 1$ in $A(n)$. This construction is illustrated in Figure 4-3 in Chapter 4. As in Chapter 4, we assign to each edge $e \in E_G$ at level ℓ in G (i.e., between nodes at levels ℓ and $\ell - 1$) a capacity

$$c_e \triangleq \begin{cases} (4^{-\ell} n)^{2-\min\{3,\alpha\}/2} & \text{if } 1 \leq \ell \leq \widetilde{L}(n), \\ 1 & \text{if } \ell = \widetilde{L}(n) + 1. \end{cases}$$

With slight abuse of notation, we let for $(u, v) = e \in E_G$

$$c_{u,v} \triangleq c_e.$$

As we shall see in the following, the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ is closely related to the following quantity:

$$\widehat{\Lambda}^{\mathsf{CA}}(n) \triangleq \left\{ \lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n} : \sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda^{\mathsf{CA}}_{U,w} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v} \ \forall S \subset V_G \right\}.$$

The region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is described by various subsets $S \subset V_G$. Each such subset can be

understood as a *cut* in the graph G. For every cut $S \subset V_G$, the sum rate

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}}$$

between nodes in S and S^c (i.e., across the cut) is bounded by the sum capacity

$$\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}$$

of edges between S and S^c. Note that we only count traffic $\lambda_{U,w}^{\mathsf{CA}}$ such that all caches U are contained in S.

The first result states that for all $\alpha > 2$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is an approximate inner bound to the caching capacity region $\Lambda^{\mathsf{CA}}(n)$.

Theorem 6.1. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1(n) \geq n^{-o(1)}$ such that*

$$b_1(n)\widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n)$$

with probability $1 - o(1)$ as $n \to \infty$.

We point out that Theorem 6.1 holds only with probability $1 - o(1)$ for different reasons in the fast and slow fading case. For fast fading, the theorem holds only for node placements that are "regular" enough. A random node placement satisfies these regularity conditions with high probability as $n \to \infty$. For slow fading, Theorem 6.1 holds under the same regularity conditions on the node placement, but moreover only holds for almost all realizations of the channel gains.

The next result provides an approximate matching outer bound to $\Lambda^{\mathsf{CA}}(n)$ for large values of path-loss exponent $\alpha > 6$.

Theorem 6.2. *Under either fast or slow fading, for any $\alpha > 6$, there exists $b_2(n) \leq n^{o(1)}$ such that*

$$\Lambda^{\mathsf{CA}}(n) \subset b_2(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$$

with probability $1 - o(1)$ as $n \to \infty$.

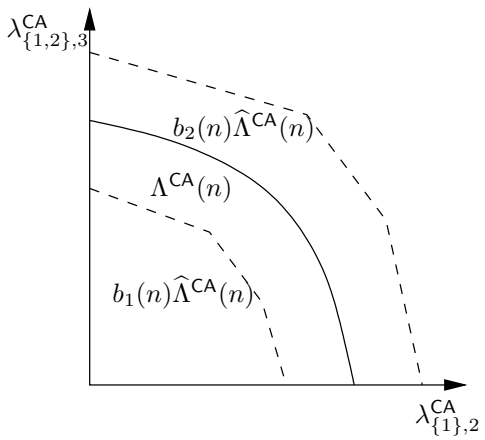

Figure 6-1: For $\alpha > 6$, the set $\widehat{\Lambda}^{\mathsf{CA}}(n)$ approximates the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ of the wireless network in the sense that $b_1(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$ (with $b_1(n) \geq n^{-o(1)}$) provides an inner bound to $\Lambda^{\mathsf{CA}}(n)$ and $b_2(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$ (with $b_2(n) \leq n^{o(1)}$) provides an outer bound to $\widehat{\Lambda}^{\mathsf{CA}}(n)$. The figure shows two dimensions (namely $\lambda^{\mathsf{UC}}_{\{1\},2}$ and $\lambda^{\mathsf{UC}}_{\{1,2\},3}$) of the $2^n \times n$-dimensional sets $\Lambda^{\mathsf{CA}}(n)$ and $\widehat{\Lambda}^{\mathsf{CA}}(n)$.

As Theorem 6.1, Theorem 6.2 holds only with high probability due to regularity conditions on the node placement. However, unlike Theorem 6.1, Theorem 6.2 holds for all realizations of channel gains also for the slow fading case.

Comparing Theorems 6.1 and 6.2, we see that for $\alpha > 6$, the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ is approximately equal to $\widehat{\Lambda}^{\mathsf{CA}}(n)$ in the sense that

$$n^{-o(1)}\widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n) \subset n^{o(1)}\widehat{\Lambda}^{\mathsf{CA}}(n).$$

In other words, for $\alpha > 6$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$ scales as the caching capacity region $\Lambda(n)$. This is illustrated in Figure 6-1.

6.1.2 Computational Aspects

As argued in previous chapters, since we are interested in large networks, computational aspects are a concern. Note that the approximate caching capacity region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is described in terms of essentially $\Theta(4^n)$ cuts $S \subset V_G$. We show in Example 6.1 in Section 6.2, that a description with significantly fewer cuts is not possible. In other words, even an approximate description $\widehat{\Lambda}^{\mathsf{CA}}(n)$ of the caching capacity region

$\Lambda^{\mathsf{CA}}(n)$ is computationally intractable for large values of n.

On the other hand, consider the simpler problem of testing membership of $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$. We now argue that this problem can be approximately solved in an efficient manner. More precisely, we show that $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can be checked approximately in polynomial time in the description complexity of λ^{CA}. Combined with Theorem 6.1 and 6.2, this shows that for $\alpha > 6$ approximate membership $\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)$ can be checked efficiently as well.

Formally, define for any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)\}.$$

Membership $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can then be evaluated by checking if $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq 1$. Let $\phi_{\lambda^{\mathsf{CA}}}(n)$ to be the solution to the following linear program

$$\begin{array}{lrll}
\max & \phi & & \\
\text{s.t.} & \sum_{p \in P_{U,w}} f_{p,U,w} & \geq \phi\lambda^{\mathsf{CA}}_{U,w} & \forall\ U \subset V(n), w \in V(n), \\
\sum_{p \in P: e \in p} \sum_{U \subset V(n)} \sum_{w \in V(n)} f_{p,U,w} & & \leq c_e & \forall\ e \in E_G, \\
& f_{p,U,w} & \geq 0 & \forall\ U \subset V(n), w \in V(n), p \in P_{U,w},
\end{array} \tag{6.1}$$

where $P_{u,w}$ is the path in G from node u to node w (since G is a tree, there is only one such paths), and where

$$P_{U,w} \triangleq \bigcup_{u \in U} P_{u,w},$$
$$P \triangleq \bigcup_{U \subset V(n)} \bigcup_{w \in V(n)} P_{U,w}.$$

Note that the linear program (6.1), and hence also $\phi_{\lambda^{\mathsf{CA}}}(n)$, can be evaluated in polynomial time in the description length of λ^{CA} (i.e., in polynomial time in the length of the "input" of the linear program) by setting the flow variables $f_{p,U,w}$ to zero whenever $\lambda^{\mathsf{CA}}_{U,w} = 0$ and $p \in P_{U,w}$. Moreover, using a primal-dual algorithm, (6.1)

can be solved efficiently in a distributed manner (see, for example, [41, Chapter 3.7]).

The following theorem shows that $\phi_{\lambda^{\mathsf{CA}}}(n)$ is a good approximation to $\rho_{\lambda^{\mathsf{CA}}}(n)$.

Theorem 6.3. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_3 \geq n^{-o(1)}$ such that for any n and caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$*

$$b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq \phi_{\lambda^{\mathsf{CA}}}(n) \leq \hat{\rho}_{\lambda^{\mathsf{CA}}}(n).$$

As argued above, $\phi_{\lambda^{\mathsf{CA}}}(n)$ can be computed in polynomial time in the description length of λ^{CA}. Hence Theorem 6.3 shows that testing membership $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can be done approximately in polynomial time in the description length of λ^{CA}. Combined with Theorems 6.1 and 6.2 this implies that, for $\alpha > 6$, approximate achievability of a traffic matrix λ^{CA} (i.e., testing membership $\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)$) can be checked efficiently and in a distributed fashion.

6.2 Example Scenarios

Here we provide three examples illustrating various aspects of the caching capacity region. Example 6.1 shows that the capacity region for caching is inherently more complicated than the ones resulting from unicast or multicast traffic. Example 6.2 shows that the strategy of always selecting the nearest cache can be arbitrarily bad. Example 6.3 analyzes the impact of complete caches on the performance of the wireless network.

Example 6.1. *Insufficiency of edge cuts*

For unicast traffic and multicast traffic, we have seen in chapters 4 and 5 that it is sufficient to consider *edge cuts* in G, i.e, cuts that result from removing a single edge from G. By construction, G has at most $2n$ edges, and hence there are at most $2n$ such edge cuts. This contrasts with the situation for caching traffic, for which Theorems 6.1 and 6.2 indicate that we have to consider general cuts, i.e., arbitrary subsets S of V_G. Indeed, the approximate capacity region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is expressed in terms of essentially $\Theta(4^n)$ cuts. Comparing these two results, one might suspect that

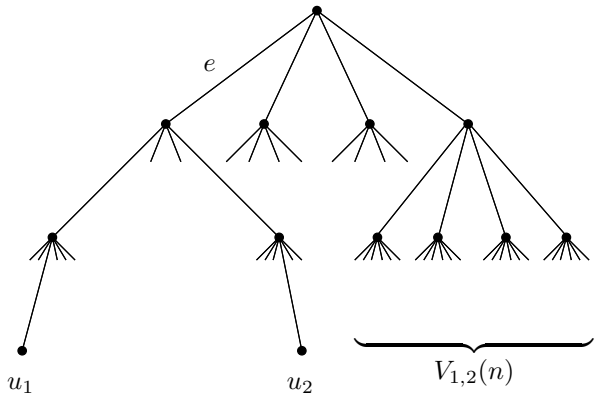

Figure 6-2: Caching traffic pattern for Example 6.1.

a simpler characterization in terms of edge cuts can be found for the caching capacity region as well. This example shows that this is not possible. In other words, the caching capacity region is inherently more complicated than the unicast or multicast capacity region of a wireless network.

Assume $V_{2,1}(n)$ and $V_{2,2}(n)$ are subsets of $V_{1,1}(n)$, and consider two nodes $u_1 \in V_{2,1}(n)$, $u_2 \in V_{2,2}(n)$. Construct

$$\lambda_{U,w}^{\mathsf{CA}} \triangleq \begin{cases} \rho(n) & \text{if } U = \{u_1, u_2\}, w \in V_{1,2}(n), \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. This is illustrated in Figure 6-2.

The best edge cut results from removing edge e in Figure 6-2. The cut capacity is $c_e = n^{2-\min\{3,\alpha\}/2}$ and the sum rate across the cut is $|V_{1,2}(n)|\rho(n)$. By Theorem 6.2 and for $\alpha > 6$, this shows that $\rho^*(n)$, the largest achievable value of $\rho(n)$, is upper bounded as

$$\rho^*(n) \leq |V_{1,2}(n)|^{-1} n^{2-\min\{3,\alpha\}/2+o(1)} = n^{1-\min\{3,\alpha\}/2+o(1)}$$

with high probability.

On the other hand, consider the general node cut $S \triangleq \{u_1, u_2\}$. The cut capacity here is 2 and the sum rate across the cut is again $|V_{1,2}(n)|\rho(n)$. Moreover, it is easily

checked that S is the bottle neck cut in G. Thus Theorem 6.1 shows that $\rho^*(n)$ is lower bounded as

$$\rho^*(n) \geq n^{-1-o(1)}, \tag{6.2}$$

and, for $\alpha > 6$, Theorem 6.2 shows that

$$\rho^*(n) \leq n^{-1-o(1)}.$$

In this example, it can be shown that the correct scaling of $\rho^*(n)$ is actually

$$\rho^*(n) = n^{-1\pm o(1)}$$

for all $\alpha > 2$ (not just $\alpha > 6$ as suggested by Theorem 6.2). Note that this differs substantially from the upper bound obtained from the best edge cut (6.2). $\Diamond$

Example 6.2. *Nearest neighbor cache selection*

A reasonable strategy of selecting caches is to request the entire message from the nearest available cache. In fact, this is the strategy implicitly assumed in most of the prior work considering caching in wireless networks cited in Section 1.2.4. This example shows that this strategy can be arbitrarily bad.

Assume $V_{2,1}(n)$ and $V_{2,2}(n)$ are subsets of $V_{1,1}(n)$, and $V_{2,3}(n)$ is a subset of $V_{1,2}(n)$. Consider a node $u^* \in V_{2,2}(n)$, and label the nodes in $V_{2,1}(n) = \{w_1, w_2, \ldots\}$ and in $V_{2,3}(n) = \{u_1, u_2, \ldots\}$. Construct

$$\lambda^{\mathsf{CA}}_{U,w} \triangleq \begin{cases} \rho(n) & \text{if } U = \{u^*, u_i\}, w = w_i \text{ for some } i, \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. This is illustrated in Figure 6-3.

For every w_i, the nearest cache is u^*. Requesting the entire message from it produces a unicast traffic pattern resulting in a per-node rate of at most

$$\rho(n) \leq n^{-1+o(1)}$$

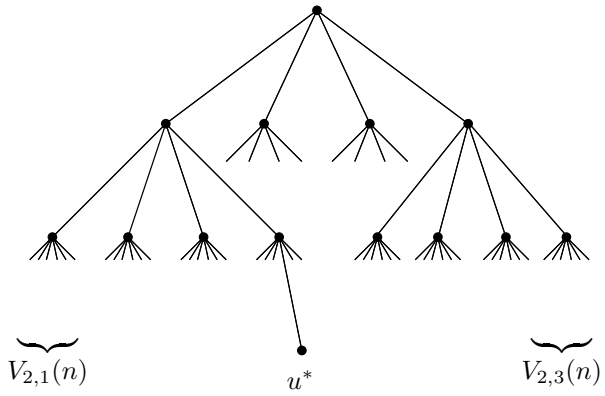

Figure 6-3: Caching traffic pattern for Example 6.2.

for all $\alpha > 2$ (see Chapter 4).

Assume now each w_i uses only the more distant cache u_i. This achieves a value of $\rho(n)$ of

$$\rho(n) \geq n^{1-\min\{3,\alpha\}/2-o(1)} \gg n^{-1+o(1)}.$$

Applying Theorem 6.1 yields the same $n^{1-\min\{3,\alpha\}/2-o(1)}$ value of $\rho(n)$, and Theorem 6.2 confirms that, for $\alpha > 6$, no scheme can achieve a better scaling. Hence

$$\rho^*(n) = n^{1-\min\{3,\alpha\}/2\pm o(1)}$$

for $\alpha > 6$, and, as in the previous example, it can be shown that this is the correct scaling of $\rho^*(n)$ also for $\alpha \in (2, 6]$. This shows that the strategy of always selecting the nearest cache can result in a scaling exponent that is considerably worse than what is achievable with optimal cache selection. $\Diamond$

Example 6.3. *Complete caches*

Assume we randomly pick n^β caches for $\beta \in [0, 1)$, each holding a complete copy of all the messages. More precisely, letting $\widetilde{W} = \{w_i\}_{i=1}^{n^\beta}$ be the collection of caches, we consider a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ of the form

$$\lambda^{\mathsf{CA}}_{W,v} = \begin{cases} \rho(n) & \text{if } W = \widetilde{W}, \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. In this setup, choosing the nearest cache strategy (as discussed in Example 6.2) results in a per-node rate of

$$\rho(n) \geq n^{\beta-1-o(1)}$$

with probability $1 - o(1)$ as $n \to \infty$. The three-layer architecture proposed in Theorem 6.1 achieves the same rate, and Theorem 6.2 shows that, for $\alpha > 6$, for any communication scheme

$$\rho(n) \leq n^{\beta-1+o(1)}.$$

Hence, for $\alpha > 6$,

$$\rho^*(n) = n^{\beta-1\pm o(1)},$$

and it can be shown, as in the previous two examples, that this is the correct scaling of $\rho^*(n)$ also for $\alpha \in (2, 6]$.

This example illustrates that when in situations in which the traffic demand and location of caches is regular enough, the strategy of selecting the nearest cache can actually be close to optimal. $\Diamond$

6.3 Communication Scheme for Caching Traffic

Theorem 6.1 provides an inner bound to the caching capacity region of a wireless network. Here we describe the communication scheme achieving this inner bound. The matching outer bound shows that, for $\alpha > 6$, this scheme is optimal in the scaling sense.

The communication architecture uses the same three layer structure introduced in 4.3 for unicast traffic. Recall that, from high to low level of abstraction, these layers are the routing layer, cooperation layer, and physical layer. Out of these three layers, only the routing layer needs to be adapted for the transmission of caching traffic.

From the view of the routing layer, the wireless network consists of the noiseless capacitated tree graph G (see Section 6.1.1 and Figure 4-3). To send a message at

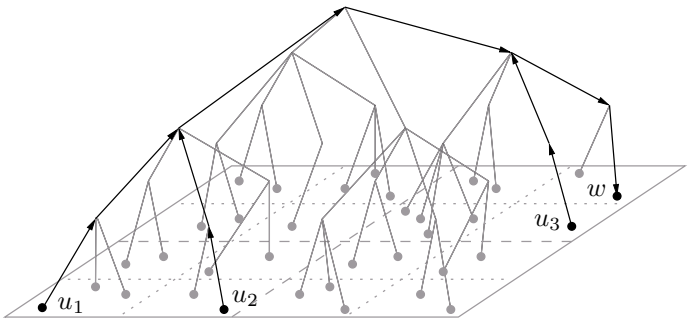

Figure 6-4: Example operation of routing layer of the three-layer architecture for caching traffic.

the caches U to its destination w, the routing layer routes the message over G. The optimal requests of message parts from the caches in U (i.e., optimal cache selection) are found by solving the linear program (6.1). As pointed out in Section 6.1.2, this optimal cache selection can be performed efficiently by a distributed algorithm. Note that this contrasts with the routing operation in the unicast and multicast cases. For unicast and multicast traffic, routing was trivial, whereas optimal routing for caching traffic is more complicated.

The next example illustrates the operations of the routing layer under caching traffic. For more details on this architecture (in particular the cooperation and physical layer), see Section 4.3.

Example 6.4. We consider a single (U, w) pair. Here, the set of caches U consist of the nodes $\{u_1, u_2, u_3\}$ in the wireless network and their destination w is in the top right of the network, as shown in Figure 6-4. Optimal cache selection is performed by solving a linear program. Note that, in general, each cache delivers only a part of the message (i.e., multiple caches are involved in communicating the message to its destination). $\Diamond$

6.4 Proofs

This section contains the proofs of Theorems 6.1, 6.2, and 6.3. We start in Section 6.4.1 with some auxiliary results. Sections 6.4.2, 6.4.3, and 6.4.4 contain the proofs

of Theorems 6.3, 6.1, and 6.2, respectively.

6.4.1 Auxiliary Results

In this section, we define some quantities needed in several of the proofs.

We first introduce a "dual" description of the various regions. Recall that for any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\left\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)\right\}.$$

Define similarly

$$\rho_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\left\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)\right\}.$$

Consider a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ for the wireless network and note that λ^{CA} can equivalently be treated as a traffic matrix between the leaf nodes of the graph G introduced in Section 6.1. Let $\Lambda_G^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$ be the collection of such caching traffic matrices $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ that can be transmitted over G using routing. Note that $\phi_{\lambda^{\mathsf{CA}}}(n)$ as defined through the linear program (6.1) is equal to

$$\phi_{\lambda^{\mathsf{CA}}}(n) = \sup\left\{\phi \geq 0 : \phi\lambda^{\mathsf{CA}} \in \Lambda_G^{\mathsf{CA}}(n)\right\}.$$

It can be shown that the regions $\Lambda^{\mathsf{CA}}(n)$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$, and $\Lambda_G^{\mathsf{CA}}(n)$ are convex, and hence knowledge of $\rho_{\lambda^{\mathsf{CA}}}(n)$, $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$, and $\phi_{\lambda^{\mathsf{CA}}}(n)$ for every $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ is sufficient to completely describe them.

Finally, recall the definition of the set $\mathcal{V}(n)$ of node placements $V(n)$ that satisfy certain regularity conditions as defined in Section 4.4.1.

6.4.2 Proof of Theorem 6.3

We first prove the upper bound, i.e.,

$$\phi_{\lambda^{\mathsf{CA}}} \leq \hat{\rho}_{\lambda^{\mathsf{CA}}}. \tag{6.3}$$

Note that if $\lambda^{\mathsf{CA}} \in \Lambda_G^{\mathsf{CA}}$ then there exists a strategy to route traffic at rates λ^{CA} over G. This implies that the flow across each cut $S \subset V_G$ must be less than the capacity of that cut. The flow across such a cut S contains at least all those requested messages that have all their caches in S and their destination in S^c, i.e.,

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}}.$$

On the other hand, the capacity of the cut S is equal to

$$\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}.$$

Thus $\lambda^{\mathsf{CA}} \in \Lambda_G^{\mathsf{CA}}$ implies

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v},$$

and hence $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}$. Therefore $\Lambda_G^{\mathsf{CA}} \subset \widehat{\Lambda}^{\mathsf{CA}}$, from which (6.3) follows.

We now prove the lower bound, i.e., we show that there exists $b_3(n) \geq n^{-o(1)}$ such that for any λ^{CA}

$$\phi_{\lambda^{\mathsf{CA}}} \geq b_3(n) \hat{\rho}_{\lambda^{\mathsf{CA}}}. \tag{6.4}$$

Pick any λ^{CA}. Since for any $b > 0$,

$$\begin{aligned} \phi_{b\lambda^{\mathsf{CA}}} &= \frac{1}{b} \phi_{\lambda^{\mathsf{CA}}}, \\ \hat{\rho}_{b\lambda^{\mathsf{CA}}} &= \frac{1}{b} \hat{\rho}_{\lambda^{\mathsf{CA}}}, \end{aligned}$$

we may assume without loss of generality that

$$\sum_{(U,w)} \lambda_{U,w}^{\mathsf{CA}} = 1. \tag{6.5}$$

Recall that G is an *undirected* capacitated graph. Construct a *directed* capacitated

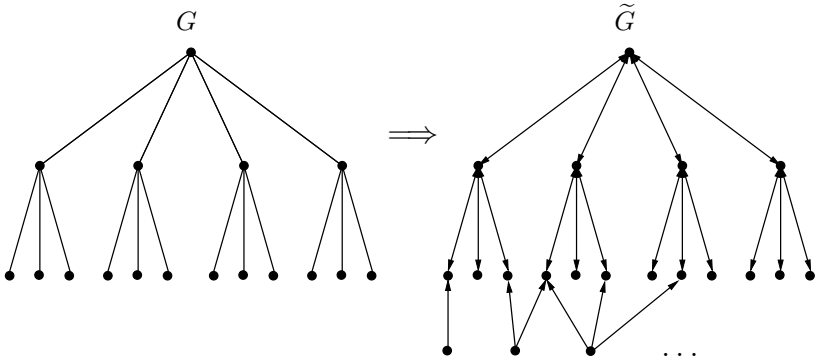

Figure 6-5: Construction of the directed graph $\widetilde{G}$ from the undirected graph G.

graph $\widetilde{G} = (V_{\widetilde{G}}, E_{\widetilde{G}})$ as follows. Take the undirected graph G and turn it into a directed graph by splitting each edge $e \in E_G$ into two directed edges each with the same capacity as e. Add 2^n additional nodes to V_G, one for each subset $U \subset V$. Connect the new node $\tilde{u}$ corresponding to U to each node $u \in U$ by a (directed) edge $(\tilde{u}, u)$ with $c_{\tilde{u},u} = \infty$. This procedure is illustrated in Figure 6-5. We call the directed version of G that is contained in $\widetilde{G}$ as a subgraph its *core*. Note that if some flows can be routed through G then the same flows can be routed through the core of $\widetilde{G}$, and if some flows can routed through the core of $\widetilde{G}$ then at least half of each flow can be routed through G. Hence, for scaling purposes, the two are equivalent.

Now, assume we are given a caching traffic matrix λ^{CA} for G. Construct a *unicast* traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ for $\widetilde{G}$ by making for each (U, w) pair in G (i.e., $U \subset V$, $w \in V$) the node $\tilde{u}$ in $\widetilde{G}$ corresponding to U a source for w with rate

$$\tilde{\lambda}^{\mathsf{UC}}_{\tilde{u},w} \triangleq \lambda^{\mathsf{CA}}_{U,w}.$$

Denote by $\Lambda^{\mathsf{UC}}_{\widetilde{G}}$ the set of feasible such unicast traffic matrices for $\widetilde{G}$, and set

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \triangleq \sup\left\{\phi \geq 0 : \phi\tilde{\lambda}^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_{\widetilde{G}}\right\}.$$

Since the edges connecting the nodes in $V_{\widetilde{G}} \setminus V_G$ to the core of $\widetilde{G}$ are in only one

direction, and by the above argument relating G to the core of $\widetilde{G}$, we have

$$\phi_{\lambda^{\mathsf{CA}}} \geq \frac{1}{2}\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}. \tag{6.6}$$

We are thus left with the problem of analyzing unicast traffic over $\widetilde{G}$. Two difficulties arise. First, $\widetilde{G}$ is a directed graph. While unicast traffic over undirected graphs with m nodes are well understood and $O(\log(m))$ approximation results for the capacity region of such graphs in terms of cut-set bounds are known [34], the best known approximation result for general directed graphs is (up to polylog factors) $O(m^{11/23})$ [2]. Second, the graph $\widetilde{G}$ is exponentially big in n. More precisely, $|V_{\widetilde{G}}| \geq 2^n$. Hence even a logarithmic (in the size m of the graph) approximation result will only yield a polynomial approximation in n. Nonetheless, as we shall see, the special structure of $\widetilde{G}$ can be exploited to obtain $\log(n)$ approximation results of $\Lambda^{\mathsf{UC}}_{\widetilde{G}}$.

We use an idea from [22], namely that the unicast traffic problem can be reduced to a maximum sum rate problem. More precisely, for a subset $\widetilde{F} \subset V_{\widetilde{G}} \times V_{\widetilde{G}}$ of (u, w) pairs in $\widetilde{G}$, define the *maximum sum rate* as

$$\tilde{\sigma}_{\widetilde{F}} \triangleq \sup\left\{\textstyle\sum_{(u,w)\in\widetilde{F}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} : \tilde{\lambda}^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_{\widetilde{G}}\right\}.$$

We now argue that for every unicast traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ there exists $\widetilde{F}$ such that $\tilde{\sigma}_{\widetilde{F}}$ is not too much bigger than $\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}$.

First, note that $\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}$ is the solution to the following linear program

$$\begin{array}{lrll} \text{maximize} & \phi & & \\ \text{subject to} & \sum_{p\in\widetilde{P}_{u,w}} f_p & \geq \phi\tilde{\lambda}^{\mathsf{UC}}_{u,w} & \forall\ u, w \in V_{\widetilde{G}}, \\ & \sum_{p\in\widetilde{P}:e\in p} f_p & \leq c_e & \forall e \in E_{\widetilde{G}}, \\ & f_p & \geq 0 & \forall\ p \in \widetilde{P}, \end{array} \tag{6.7}$$

where $\widetilde{P}_{u,w}$ is the collection of all paths in $\widetilde{G}$ from node u to node w, and

$$\widetilde{P} \triangleq \bigcup_{(u,w)\in V_{\widetilde{G}}\times V_{\widetilde{G}}} \widetilde{P}_{u,w}.$$

The corresponding dual linear program is

$$\begin{array}{lrll} \text{minimize} & \sum_{e\in E_{\widetilde{G}}} c_e m_e & & \\ \text{subject to} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall\ u,w \in V_{\widetilde{G}}, p \in \widetilde{P}_{u,w}, \\ & \sum_{u,w\in V_{\widetilde{G}}} d_{u,w}\tilde{\lambda}^{\mathsf{UC}}_{u,w} & \geq 1 & \\ & m_e & \geq 0 & \forall\ e \in E_{\widetilde{G}}, \\ & d_{u,w} & \geq 0 & \forall\ u,w \in V_{\widetilde{G}}. \end{array} \tag{6.8}$$

Since the all-zero solution is feasible for the primal program (6.7), strong duality holds.

Second, $\tilde{\sigma}_{\widetilde{F}}$ is the solution to the linear program

$$\begin{array}{lrll} \text{maximize} & \sum_{(u,w)\in\widetilde{F}} \sum_{p\in\widetilde{P}_{u,w}} f_p & & \\ \text{subject to} & \sum_{p\in\widetilde{P}:e\in p} f_p & \leq c_e & \forall e \in E_{\widetilde{G}}, \\ & f_p & \geq 0 & \forall\ p \in \widetilde{P}, \end{array}$$

and its dual is

$$\begin{array}{lrll} \text{minimize} & \sum_{e\in E_{\widetilde{G}}} c_e m_e & & \\ \text{subject to} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall\ u,w \in V_{\widetilde{G}}, p \in \widetilde{P}_{u,w}, \\ & d_{u,w} & \geq 1 & \forall\ (u,w) \in \widetilde{F}, \\ & m_e & \geq 0 & \forall\ e \in E_{\widetilde{G}}, \\ & d_{u,w} & \geq 0 & \forall\ u,w \in V_{\widetilde{G}}. \end{array} \tag{6.9}$$

Again strong duality holds.

Let $\{m^*_e\}_{e\in E_{\widetilde{G}}}$, $\{d^*_{u,w}\}_{u,w\in V_{\widetilde{G}}}$ be a minimizer for the dual (6.8) of the unicast traffic problem. We now show how $\{m^*_e\}$, $\{d^*_{u,w}\}$ can be used to construct a solution to the dual (6.9) of the maximum sum rate problem. Note that we can assume without loss

of optimality that

$$d^*_{u,w} = \begin{cases} 0 & \text{if } \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 0, \\ \min_{p \in \widetilde{P}_{u,w}} \sum_{e \in p} m^*_e & \text{else.} \end{cases} \tag{6.10}$$

Now, since $c_e = \infty$ whenever $e \in E_{\widetilde{G}} \setminus E_G$, we have $m^*_e = 0$ for those edges. Since in addition $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$ only if $u \in V_{\widetilde{G}} \setminus V_G$, this implies that $\{d^*_{u,w}\}_{u,w \in V_{\widetilde{G}}}$ can take at most n^2 different nonzero values. Order these values in decreasing order

$$d^*_1 > d^*_2 > \ldots > d^*_K > d^*_{K+1} = 0$$

with $K \leq n^2$, and define

$$\tilde{\lambda}^{\mathsf{UC}}_k \triangleq \sum_{u,w \in V_{\widetilde{G}} : d^*_{u,w} = d^*_k} \tilde{\lambda}^{\mathsf{UC}}_{u,w}.$$

We now argue that $d^*_k \leq n^2$ for all $k \in \{1, \ldots, K\}$. In fact, assume $d^*_1 > n^2$, then by (6.10) there exists at least one edge $\tilde{e}$ such that $m^*_{\tilde{e}} > n$. Hence

$$\sum_{e \in E_{\widetilde{G}}} c_e m^*_e \geq c_{\tilde{e}} m^*_{\tilde{e}} > n$$

since $c_e \geq 1$ for all $e \in E_{\widetilde{G}}$. On the other hand, let $m_e = 1$ for all edges between the leave nodes and parent nodes in the core of $\widetilde{G}$, and let $m_e = 0$ for all other edges. Set $d_{u,w}$ as in (6.10) but with respect to $\{m_e\}$. Since all paths between node pairs (u, w) such that $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$ include at least one edge between the aforementioned leave and parent nodes, we have

$$\sum_{u,w \in V_{\widetilde{G}}} d_{u,w} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \geq \sum_{u,w \in V_{\widetilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 1,$$

by the normalization assumption (6.5). Thus $\{m_e\}$, $\{d_{u,v}\}$ is feasible for the dual (6.8), and has value

$$\sum_{e \in E_{\widetilde{G}}} c_e m_e = n < \sum_{e \in E_{\widetilde{G}}} c_e m^*_e,$$

contradicting the optimality of $\{m_e^*\}$, $\{d_{u,v}^*\}$. Hence $d_k^* \le d_1^* \le n^2$ for all k.

We now argue that at least one d_k^* is not too small. Let $k_1 < k_2 < \ldots < k_I$ be such that

$$\{k_i\}_{i=1}^I = \Big\{k : \tilde{\lambda}_k^{\mathsf{UC}} \ge \frac{1}{2n^4}\Big\}. \tag{6.11}$$

Note that $I \ge 1$ since otherwise

$$\begin{aligned}\sum_{u,w \in V_{\widetilde{G}}} \tilde{\lambda}_{u,w}^{\mathsf{UC}} &= \sum_{k=1}^{K+1} \tilde{\lambda}_k^{\mathsf{UC}} \\ &< (K+1)\frac{1}{2n^4} \\ &\le \frac{n^2+1}{2n^4} \\ &\le 1,\end{aligned}$$

contradicting the normalization assumption (6.5). Finally, define for $i \le I$,

$$s_i \triangleq \sum_{j=1}^{i} \tilde{\lambda}_{k_j}^{\mathsf{UC}}.$$

Using that $\{d_k^*\}$ is feasible for the dual (6.8), that $d_k^* \le n^2$, and that $K \le n^2$, we have

$$\begin{aligned}\sum_{i=1}^{I} d_{k_i}^* \tilde{\lambda}_{k_i}^{\mathsf{UC}} &\ge 1 - \sum_{k:\tilde{\lambda}_k^{\mathsf{UC}} < 1/2n^4} d_k^* \tilde{\lambda}_k^{\mathsf{UC}} \\ &\ge 1 - \frac{1}{2n^4} K n^2 \\ &\ge \frac{1}{2}.\end{aligned} \tag{6.12}$$

We argue that this implies existence of i such that

$$d_{k_i}^* \ge \frac{1}{2s_i(1+\ln(2n^4))}. \tag{6.13}$$

Indeed, assume (6.13) is false for all i. Then

$$\begin{aligned}
\sum_{i=1}^{I} d^*_{k_i}\tilde{\lambda}^{\mathsf{UC}}_{k_i} &< \frac{1}{2(1+\ln(2n^4))}\sum_{i=1}^{I}\frac{\tilde{\lambda}^{\mathsf{UC}}_{k_i}}{s_i} \\
&= \frac{1}{2(1+\ln(2n^4))}\Big(1+\sum_{i=2}^{I}\frac{s_i-s_{i-1}}{s_i}\Big) && \text{(6.14a)} \\
&\le \frac{1}{2(1+\ln(2n^4))}\Big(1+\sum_{i=2}^{I}\big(\ln(s_i)-\ln(s_{i-1})\big)\Big) && \text{(6.14b)} \\
&= \frac{1}{2(1+\ln(2n^4))}\big(1+\ln(s_I/\tilde{\lambda}^{\mathsf{UC}}_{k_1})\big)\big) \\
&\le \frac{1}{2(1+\ln(2n^4))}\big(1+\ln(2n^4)\big) && \text{(6.14c)} \\
&= \frac{1}{2},
\end{aligned}$$

where we have used that $I \ge 1$ in (6.14a), that $1-x \le -\ln(x)$ for every $x \ge 0$ in (6.14b), and that $s_I \le 1$ by (6.5) and $\tilde{\lambda}^{\mathsf{UC}}_{k_1} \ge \frac{1}{2n^4}$ in (6.14c). This contradicts (6.12), showing (6.13) must hold for some i. Consider this value of i in the following.

Now, consider the following set $\widetilde{F}$ of (u,w) pairs:

$$\widetilde{F} \triangleq \big\{(u,w) : d^*_{u,w} \ge d^*_{k_i}\big\}.$$

Note that by (6.10) we have that if $(u,w) \in \widetilde{F}$ then $u \in V_{\widetilde{G}} \setminus V_G$ and w is a "leaf" node in V_G. In other words, (u,w) pairs in $\widetilde{F}$ correspond to caches-destination pairs (U,w) in G. Set for all $u,w \in V_{\widetilde{G}}$ and $e \in E_{\widetilde{G}}$,

$$\begin{aligned}
d_{u,w} &\triangleq \frac{d^*_{u,w}}{d^*_{k_i}}, \\
m_e &\triangleq \frac{m^*_e}{d^*_{k_i}}.
\end{aligned}$$

Note that for $(u,w) \in \widetilde{F}$,

$$d_{u,w} = \frac{d^*_{u,w}}{d^*_{k_i}} \ge 1,$$

and that for all $u, w \in V_{\widetilde{G}}$, $p \in \widetilde{P}_{u,w}$,

$$\begin{aligned}\sum_{e\in p} m_e &= \frac{1}{d^*_{k_i}} \sum_{e\in p} m^*_e \\ &\geq \frac{1}{d^*_{k_i}} d^*_{u,w} \\ &= d_{u,w},\end{aligned}$$

by feasibility of $\{d^*_{u,w}\}$ and $\{m^*_e\}$ for the dual (6.8). Hence, for this $\widetilde{F}$, the choice of $\{m_e\}$ and $\{d_{u,w}\}$ is feasible for the dual (6.9).

By weak duality

$$\begin{aligned}\tilde{\sigma}_{\widetilde{F}} &\leq \sum_{e\in E_{\widetilde{G}}} c_e m_e \\ &= \frac{1}{d^*_{k_i}} \sum_{e\in E_{\widetilde{G}}} c_e m^*_e.\end{aligned}$$

By (6.13),

$$d^*_{k_i} \geq \frac{1}{2s_i(1+\ln(2n^4))},$$

and

$$\begin{aligned}s_i &= \sum_{j=1}^{i} \tilde{\lambda}^{\mathsf{UC}}_{k_j} \\ &= \sum_{j=1}^{i} \sum_{(u,w):d^*_{u,w}=d^*_{k_j}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &\leq \sum_{(u,w):d^*_{u,w}\geq d^*_{k_i}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &= \sum_{(u,w)\in\widetilde{F}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &\triangleq \tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}},\end{aligned}$$

where in the third line we have used that $d^*_{k_1} > d^*_{k_2} > \ldots$. Therefore

$$\tilde{\sigma}_{\widetilde{F}} \leq 2\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}(1+\ln(2n^4)) \sum_{e\in E_{\widetilde{G}}} c_e m^*_e.$$

Since $\{m^*_e\}$ is optimal for the dual (6.8), and by strong duality, we also have

$$\sum_{e\in E_{\widetilde{G}}} c_e m^*_e = \tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}},$$

and hence

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \geq \frac{1}{2(1+\ln(2n^4))} \frac{\tilde{\sigma}_{\widetilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}}. \tag{6.15}$$

We are thus left with analyzing maximum sum rates $\tilde{\sigma}_{\widetilde{F}}$. Now notice that since the edges in $E_{\widetilde{G}} \setminus E_G$ have infinite capacity, and since for $(u,w) \in \widetilde{F}$ we can assume without loss of generality that $u \in V_{\widetilde{G}} \setminus V_G$, this analysis can be done by considering only the core of $\widetilde{G}$. More precisely, for a collection of node pairs $\widetilde{F}$ in $\widetilde{G}$ as above, we construct a collection of node pairs F in G as follows. For each $(u,w) \in \widetilde{F}$ with $\tilde{u}$ connected by $\widetilde{G}$ with nodes $U \subset V_G \subset V_{\widetilde{G}}$, add (u,w) to F for each $u \in U$. Denote by σ_F the maximum sum rate for F in G. Since G is the undirected version of the core of $\widetilde{G}$, we have

$$\tilde{\sigma}_{\widetilde{F}} \geq \sigma_F. \tag{6.16}$$

For a collection of node pairs F in G, we call a set of edges M a *multicut* for F if in the graph $(V_G, E_G \setminus M)$ each pair in F is disconnected. For a subset $M \subset E_G$, define

$$c_M \triangleq \sum_{e\in M} c_e.$$

It is shown in [13, Theorem 8] that if G is an undirected tree, then for every $F \in V_G \times V_G$ there exists a multicut M for F such that

$$\sigma_F \geq \frac{1}{2} c_M. \tag{6.17}$$

Combining (6.15), (6.16), and (6.17), we obtain that for every $\tilde{\lambda}^{\mathsf{UC}}$ there exists a

collection of node pairs $\widetilde{F}$ in $\widetilde{G}$, and a multicut M for the corresponding F in G such that

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \geq \frac{1}{4(1+\ln(2n^4))} \frac{c_M}{\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}}. \tag{6.18}$$

We now show how the edge cut $M \subset E_G$ can be transformed into a node cut $S \subset V_G$. Denote by $\{S_i\}$ the connected components of $(V_G, E_G \setminus M)$. We have

$$\sum_i c_{(S_i^c \times S_i) \cap E_G} \leq 2c_M, \tag{6.19}$$

since $(S_i \times S_i^c) \cap E_G \subset M$ for every i, and since every edge in M appears at most twice in the sum on the left-hand side. With slight abuse of notation, define for $S \subset V_G$

$$\lambda^{\mathsf{CA}}_{S,S^c} \triangleq \sum_{U \subset S \cap V} \sum_{w \in V \setminus S} \lambda^{\mathsf{CA}}_{U,w}.$$

M is a multicut for the F induced by $\widetilde{F}$, and hence for every $(u,w) \in \widetilde{F}$ and the corresponding pair (U,w), M separates w from all the nodes in U. Therefore, for each such (U,w) pair, there exists a S_i such that $w \in S_i$, $U \cap S_i = \emptyset$. This shows that

$$\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}} \leq \sum_i \lambda^{\mathsf{CA}}_{S_i^c, S_i}. \tag{6.20}$$

Equations (6.18), (6.19), and (6.20) imply that there exists j such that

$$\begin{aligned}
\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} &\geq \frac{1}{8(1+\ln(2n^4))} \frac{\sum_i c_{(S_i^c \times S_i) \cap E_G}}{\sum_i \lambda^{\mathsf{CA}}_{S_i^c, S_i}} \\
&\geq \frac{1}{8(1+\ln(2n^4))} \frac{c_{(S_j^c \times S_j) \cap E_G}}{\lambda^{\mathsf{CA}}_{S_j^c, S_j}} \\
&\geq \frac{1}{8(1+\ln(2n^4))} \min_{S \subset V_G} \frac{c_{(S \times S^c) \cap E_G}}{\lambda^{\mathsf{CA}}_{S,S^c}} \\
&= \frac{1}{8(1+\ln(2n^4))} \hat{\rho}_{\lambda^{\mathsf{CA}}}.
\end{aligned} \tag{6.21}$$

Combined with (6.6), this shows that for

$$b_3(n) \triangleq \frac{1}{16(1+\ln(2n^4))} \geq n^{-o(1)}$$

we have

$$\phi_{\lambda^{\mathsf{CA}}} \geq b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}},$$

proving the lower bound in Theorem 6.3.

6.4.3 Proof of Theorem 6.1

In this Section, we provide the proof of Theorem 6.1. Instead of proving the theorem directly, it will be convenient to work with the dual description $\rho_{\lambda^{\mathsf{CA}}}(n)$ and $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$ introduced in Section 6.4.1. The next theorem proves the dual version of Theorem 6.1.

Theorem 6.4. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1 = n^{-o(1)}$ such that with probability $1 - o(1)$ as $n \to \infty$ for any n and any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$*

$$b_1(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq \rho_{\lambda^{\mathsf{CA}}}(n).$$

Proof. The same construction as in the proof of Theorem 4.10 shows that there exists $b(n) \geq n^{-o(1)}$ such that if a caching traffic matrix λ^{CA} can be routed over G, then $b(n)\lambda^{\mathsf{CA}}$ can be communicated reliably over the wireless network. Formally, if $V \in \mathcal{V}$ then under fast fading

$$b(n)\phi_{\lambda^{\mathsf{CA}}} \leq \rho_{\lambda^{\mathsf{CA}}}, \tag{6.22}$$

and the same results holds for slow fading for a collection of channel gains $\mathcal{H}$ (not dependent on λ^{CA}) with

$$\mathbb{P}(\{h_{u,v}\}_{u,v\in V} \in \mathcal{H}) \geq 1 - o(1)$$

as $n \to \infty$.

Combining (6.22), with Theorem 6.3 and Lemma 4.5, we obtain that with probability

$$\mathbb{P}(\{h_{u,v}\} \in \mathcal{H}, V \in \mathcal{V}) \geq 1 - o(1)$$

as $n \to \infty$, we have for any caching traffic matrix λ^{CA}

$$\begin{aligned}\rho_{\lambda^{\mathsf{CA}}} &\geq b(n)\phi_{\lambda^{\mathsf{CA}}} \\ &\geq b(n)b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}.\end{aligned}$$

Setting

$$b_1(n) \triangleq b(n)b_3(n),$$

and recalling that $b_3(n) \geq n^{-o(1)}$ and $b(n) \geq n^{-o(1)}$ both uniformly in λ^{CA}, concludes the proof of Theorem 6.4. □

6.4.4 Proof of Theorem 6.2

In this Section, we prove Theorem 6.2. As before, it will be convenient to work with the dual description $\rho_{\lambda^{\mathsf{CA}}}(n)$ and $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$ introduced in Section 6.4.1. The next theorem proves the dual version of Theorem 6.2

Theorem 6.5. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_2 \leq n^{o(1)}$ such that with probability $1 - o(1)$ as $n \to \infty$ for any n and any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$*

$$\rho_{\lambda^{\mathsf{CA}}}(n) \leq b_2(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n).$$

We start with some auxiliary lemmas. For a subsets $S_1, S_2 \subset V(n)$, denote by $C(S_1, S_2)$ the MIMO capacity between the nodes in S_1 and S_2. Moreover, denote by S_2^k the nodes in S_2 that are at distance between k and $k+1$ from S_1, i.e.,

$$S_2^k \triangleq \{v \in S_2 : \min_{u \in S_1} r_{u,v} \in [k, k+1)\}.$$

Lemma 6.6. *Under either fast or slow fading, for every $\alpha > 6$, there exists a constant*

K_1 such that for all $V(n) \in \mathcal{V}(n)$ and all $S_1, S_2 \subset V(n)$, $S_2 \cap S_1 = \emptyset$,

$$C(S_1, S_2) \leq K_1 \log^4(n) \sum_{k=0}^{\log(n)} |S_2^k|.$$

Proof. Note that

$$S_2 = \bigcup_{k=0}^{\infty} S_2^k,$$

Let

$$\boldsymbol{H}_{S_1,S_2} \triangleq [h_{u,v}]_{u \in S_1, v \in S_2}$$

be the matrix of channel gains between the nodes in S_1 and S_2. Under fast fading

$$C(S_1, S_2) \triangleq \max_{\substack{\boldsymbol{Q}(\boldsymbol{H}) \geq 0: \\ \mathbb{E}(q_{u,u}) \leq P \ \forall u \in S_1}} \mathbb{E}\bigg(\log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^{\dagger} \boldsymbol{Q}(\boldsymbol{H}) \boldsymbol{H}_{S_1,S_2}\right)\bigg),$$

and under slow fading

$$C(S, S^c) \triangleq \max_{\substack{\boldsymbol{Q} \geq 0: \\ q_{u,u} \leq P \ \forall u \in S_1}} \log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^{\dagger} \boldsymbol{Q} \boldsymbol{H}_{S_1,S_2}\right).$$

Applying the generalized Hadamard inequality, we obtain that under either fast or slow fading

$$C(S_1, S_2) \leq C(S, \cup_{k=0}^{\log(n)} S_2^k) + C(S, \cup_{k > \log(n)} S_2^k). \tag{6.23}$$

For the first term in (6.23), using Hadamard's inequality once more, yields

$$\begin{aligned} C(S_1, \cup_{k=0}^{\log(n)} S_2^k) &\leq \sum_{k=0}^{\log(n)} \sum_{v \in S_2^k} C(S_1, \{v\}) \\ &\leq \sum_{k=0}^{\log(n)} \sum_{v \in S_2^k} C(\{v\}^c, \{v\}). \end{aligned}$$

By Lemma 4.6,

$$C(\{v\}^c, \{v\}) \leq K \log(n),$$

and thus

$$C(S_1, \cup_{k=0}^{\log(n)} S_2^k) \leq K \log(n) \sum_{k=0}^{\log(n)} |S_2^k|. \tag{6.24}$$

For the second term in (6.23), we have the following upper bound from (slightly adapting) Theorem 2.1 in [21]:

$$C(S_1, \cup_{k>\log(n)} S_2^k) \leq \sum_{k>\log(n)} \sum_{v \in S_2^k} \Big(\sum_{u \in S_1} r_{u,v}^{-\alpha/2} \Big)^2.$$

Consider $v \in S_2^k$. By the definition of S_2^k, the (open) ball of radius k around v does not contain any node in S_1. Moreover, since $V \in \mathcal{V}$, there are at most $\log(n)$ nodes inside every square of sidelength one. Thus

$$\begin{aligned} \sum_{u \in S_1} r_{u,v}^{-\alpha/2} &\leq 4\pi(k+2)^2 \log(n) k^{-\alpha/2} + \log(n) \sum_{\tilde{k}=2k}^{\infty} 10\pi(\tilde{k}+2)\tilde{k}^{-\alpha/2} \\ &\leq \widetilde{K} \log(n) k^{2-\alpha/2}, \end{aligned}$$

for some constant K independent of S_1 and k. Therefore,

$$C(S_1, \cup_{k>\log(n)} S_2^k) \leq \sum_{k>\log(n)} |S_2^k| \widetilde{K}^2 \log^2(n) k^{4-\alpha}. \tag{6.25}$$

Consider now some $v \in S_2^k$ with $k > \log(n)$, and let u^* be the closest point in S_1 to v. Since $v \in S_2^k$, we must have

$$r_{u^*,v} \in [k, k+1).$$

Consider the (open) ball of radius $r_{u^*,v}$ around v and the ball of radius $\log(n)$ around u^*. Since u^* is the closest node to v in S_1, all nodes in the ball around v are in S_2. Moreover, the intersection of the two balls has an area of at least $\frac{\pi}{4} \log^2(n)$. Since $V \in \mathcal{V}$, this implies that this intersection must contain at least one point, say $\tilde{v}$, and

by construction

$$\tilde{v} \in \bigcup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}.$$

This shows that for every node v in S_2^k there exists a node $\tilde{v}$ in $\cup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}$ such that

$$r_{v,\tilde{v}} \in [k - \log(n), k + 1).$$

Now, since $V \in \mathcal{V}$, for every node $\tilde{v}$, there are at most

$$2\pi(k+1)(\log(n)+6)\log(n) \leq K' k \log^2(n)$$

nodes at distance $[k-\log(n), k+1)$ for some constant K'. Hence the number of nodes in S_2^k is at most

$$|S_2^k| \leq K' k \log^2(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|. \tag{6.26}$$

Combining (6.26) with (6.25) yields

$$\begin{aligned} C(S_1, \cup_{k>\log(n)} S_2^k) &\leq \sum_{k>\log(n)} |S_2^k| \widetilde{K}^2 \log^2(n) k^{4-\alpha} \\ &\leq K' \widetilde{K}^2 \log^4(n) \Big(\sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}| \Big) \sum_{k>\log(n)} k^{5-\alpha} \\ &= K'' \log^4(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|, \end{aligned} \tag{6.27}$$

for some constant K'', and where we have used that $\alpha > 6$. Finally, substituting (6.24) and (6.27) into (6.23) shows that

$$C(S_1, S_2) \leq (K + K'') \log^4(n) \sum_{k=0}^{\log(n)} |S_2^k|,$$

which proves the lemma with

$$K_1 \triangleq K + K''. \qquad \square$$

The next lemma shows that for large path-loss exponents ($\alpha > 6$) every cut is approximately achievable, i.e., for every cut there exists an achievable unicast traffic matrix that has a sum rate across the cut that is not much smaller than the cut capacity.

Lemma 6.7. *Under fast fading, for every $\alpha > 6$, there exists $b_4(n) \leq n^{o(1)}$ such that for $V(n) \in \mathcal{V}(n)$ and $S \subset V(n)$ we can find a unicast traffic matrix $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$ satisfying*

$$C(S, S^c) \leq b_4(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w}. \tag{6.28}$$

Moreover, there exists a collection of channel gains $\mathcal{H}(n)$ such that

$$\mathbb{P}\big(\{h_{u,v}\}_{u,v \in V(n)} \in \mathcal{H}(n)\big) \geq 1 - o(1)$$

as $n \to \infty$, and such that for $\{h_{u,v}\}_{u,v} \in \mathcal{H}(n)$, (6.28) holds for slow fading as well.

Proof. By Lemma 6.6 for $V \in \mathcal{V}$

$$C(S, S^c) \leq K_1 \log^4(n) |\{v \in S^c : r_{S,v} < \log(n) + 1\}|, \tag{6.29}$$

where

$$r_{S,v} \triangleq \min_{u \in S} r_{u,v}.$$

Construct a unicast traffic matrix $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$ as

$$\lambda^{\mathsf{UC}}_{u,w} \triangleq \begin{cases} \rho(n) & \text{if } r_{u,w} < \log(n) + 1, \\ 0 & \text{else,} \end{cases}$$

for some function $\rho(n)$. We now argue that for $\rho(n) = \Theta(\log^{-2}(n))$ there exists $\tilde{b}(n) \geq n^{-o(1)}$ such that $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$. This follows from Theorem 4.1, once we show that for every $\ell \in \{1, \ldots, \widetilde{L}(n)\} \cup \{\log(n)\}$ and $i \in \{1, \ldots, 4^\ell\}$ we have

$$\sum_{u \in V_{\ell,i}} \sum_{w \notin V_{\ell,i}} \lambda^{\mathsf{UC}}_{u,w} \leq \max\{1, 4^{-\ell} n\}^{2 - \min\{3, \alpha\}/2},$$

$$\sum_{u \notin V_{\ell,i}} \sum_{w \in V_{\ell,i}} \lambda^{\mathsf{UC}}_{u,w} \leq \max\{1, 4^{-\ell} n\}^{2-\min\{3,\alpha\}/2},$$

and

$$\sum_{u \neq w} \lambda^{\mathsf{UC}}_{u,w} \leq K \log^2(n) \rho(n) \qquad \forall w \in V,$$
$$\sum_{w \neq u} \lambda^{\mathsf{UC}}_{u,w} \leq K \log^2(n) \rho(n) \qquad \forall u \in V,$$

for some constant K. By the locality of the traffic matrix λ^{UC}, this is sufficient to show that Theorem 4.1 applies for $\rho(n) = \frac{1}{K} \log^{-2}(n)$. Hence $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ for fast fading, and the same conclusion holds for slow fading for some $\mathcal{H}$ with

$$\mathbb{P}\big(\{h_{u,v}\}_{u,v \in V} \in \mathcal{H}\big) \geq 1 - o(1)$$

as $n \to \infty$.

Combined with (6.29), this implies that

$$C(S, S^c) \leq \frac{K \log^6(n)}{\tilde{b}(n)} \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w},$$

proving the lemma. □

We are now ready for the proof of the outer bound on $\Lambda^{\mathsf{CA}}(n)$.

Proof of Theorem 6.5. Consider a cut $S \subset V$ in the wireless network. Assume we allow the nodes on each side of the cut to cooperate without any restriction — this can clearly only increase $\rho_{\lambda^{\mathsf{CA}}}$. The total amount of traffic that needs to be transmitted across the cut is then

$$\sum_{W \subset S} \sum_{v \notin S} \lambda^{\mathsf{CA}}_{W,v}.$$

The maximum achievable sum rate (with the aforementioned node cooperation) is

given by $C(S, S^c)$ the MIMO capacity between the nodes in S and in S^c. Therefore

$$\rho_{\lambda^{\mathsf{CA}}} \leq \min_{S \subset V} \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda^{\mathsf{CA}}_{U,w}}. \tag{6.30}$$

We proceed by relating the cut S in the wireless network to a cut $\widetilde{S}$ in G. By Lemma 6.7 for $V \in \mathcal{V}$, there exists $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ such that for fast fading

$$C(S, S^c) \leq b_4(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w}, \tag{6.31}$$

and (6.31) holds also for slow fading if $\{h_{u,v}\}_{u,v} \in \mathcal{H}$ (with $\mathcal{H}$ defined as in Lemma 6.7). By Theorem 4.2, for $\alpha > 5$ and $V \in \mathcal{V}$ there exists K such that if $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ then $K \log^{-6}(n) \lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_G$.

Now, consider any $\widetilde{S} \subset V_G$ such that $\widetilde{S} \cap V = S$. Note that $\widetilde{S}$ is a cut in G separating S from $V \setminus S$. Since $K \log^{-6}(n) \lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_G$, we thus have

$$\sum_{u \in S} \sum_{w \notin S} K \log^{-6}(n) \lambda^{\mathsf{UC}}_{u,w} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in \widetilde{S}, v \notin \widetilde{S}}} c_{u,v},$$

and by minimizing over the choice of $\widetilde{S}$,

$$\sum_{u \in S} \sum_{w \notin S} K \log^{-6}(n) \lambda^{\mathsf{UC}}_{u,w} \leq \min_{\widetilde{S}: \widetilde{S} \cap V = S} \sum_{\substack{(u,v) \in E_G: \\ u \in \widetilde{S}, v \notin \widetilde{S}}} c_{u,v}. \tag{6.32}$$

Combining (6.31) and (6.32) shows that

$$C(S, S^c) \leq \frac{b_4(n)}{K} \log^6(n) \min_{\widetilde{S}: \widetilde{S} \cap V = S} \sum_{\substack{(u,v) \in E_G: \\ u \in \widetilde{S}, v \notin \widetilde{S}}} c_{u,v}.$$

Together with (6.30), and using Lemmas 4.5 and 6.7, this yields that with probability

$$\mathbb{P}(\{h_{u,v}\}_{u,v} \in \mathcal{H}, V \in \mathcal{V}) \geq 1 - o(1)$$

as $n \to \infty$, we have for any caching traffic matrix λ^{CA}

$$\begin{aligned}
\rho_{\lambda^{\mathsf{CA}}} &\leq \min_{S \subset V} \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda^{\mathsf{CA}}_{U,w}} \\
&\leq b_2(n) \min_{S \subset V} \min_{\widetilde{S} \in V_G : \widetilde{S} \cap V = S} \frac{\sum_{\substack{(u,v) \in E_G : \\ u \in \widetilde{S}, v \notin \widetilde{S}}} c_{u,v}}{\sum_{U \subset \widetilde{S} \cap V} \sum_{w \in V \setminus \widetilde{S}} \lambda^{\mathsf{CA}}_{U,w}} \\
&= b_2(n) \min_{\widetilde{S} \subset V_G} \frac{\sum_{\substack{(u,v) \in E_G : \\ u \in \widetilde{S}, v \notin \widetilde{S}}} c_{u,v}}{\sum_{U \subset \widetilde{S} \cap V} \sum_{w \in V \setminus \widetilde{S}} \lambda^{\mathsf{CA}}_{U,w}} \\
&= b_2(n) \hat{\rho}_{\lambda^{\mathsf{CA}}},
\end{aligned}$$

with

$$b_2(n) \triangleq \frac{b_4(n)}{K} \log^6(n) \leq n^{o(1)}.$$

□

6.5 Discussion

Here we discuss extensions and implications of the results presented in this chapter. In Section 6.5.1 we consider dense networks, and in Section 6.5.2 we discuss design guidelines.

6.5.1 Dense Networks

The results presented so far in this chapter assumed extended node placement. Here we discuss how these results can be modified for the dense case, where n nodes are located on a square of area one.

As was the case for unicast and multicast traffic (see chapters 4 and 5), achievability for dense networks (regardless of the value of path-loss exponent) can be derived from the achievability result for extended networks by taking a limit as $\alpha \to 2$. This shows that for dense networks, for any $\alpha > 2$, there exists $b_1(n) \geq n^{-o(1)}$ such that with probability $1 - o(1)$

$$b_1(n) \widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n)$$

with

$$\widehat{\Lambda}^{\mathsf{CA}}(n) \triangleq \left\{\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n} : \sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v} \ \forall S \subset V_G\right\},$$

and for an edge e at level ℓ in G,

$$c_e \triangleq \begin{cases} 4^{-\ell} n & \text{if } 1 \leq \ell \leq \widetilde{L}(n), \\ 1 & \text{if } \ell = \widetilde{L}(n) + 1. \end{cases}$$

For the converse result, we need to replace Lemma 6.6 by the inequality

$$C(S, S^c) \leq K \log(n) \min\{|S|, |S^c|\}$$

for some constant K, and we need to modify Lemma 6.7 accordingly. From this, we obtain that for dense networks, for any $\alpha > 2$, there exists $b_2(n) \leq n^{o(1)}$ such that with probability $1 - o(1)$

$$\Lambda^{\mathsf{CA}}(n) \subset b_2(n) \widehat{\Lambda}^{\mathsf{CA}}(n).$$

This is a stronger result than the corresponding upper bound for extended networks, which was only shown to hold for $\alpha > 6$.

In other words, for dense networks, for any value of path-loss exponent $\alpha > 2$, we have $\Lambda^{\mathsf{CA}}(n) = n^{\pm o(1)} \widehat{\Lambda}^{\mathsf{CA}}(n)$. Furthermore, in the dense setting, the expression for $\widehat{\Lambda}^{\mathsf{CA}}(n)$ can be simplified to

$$\widehat{\Lambda}^{\mathsf{CA}}(n) = \left\{\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n} : \sum_{U \subset S} \sum_{w \in S^c} \lambda_{U,w}^{\mathsf{CA}} \leq \min\{|S|, n - |S|\} \ \forall S \subset V(n)\right\}.$$

Note that here we only consider subsets of $V(n)$ (as opposed to $V_G(n)$). Hence, for the dense case $\widehat{\Lambda}^{\mathsf{CA}}(n)$ provides a complete scaling characterization of the $2^n \times n$-dimensional caching capacity region $\Lambda^{\mathsf{CA}}(n)$ in terms of $\Theta(2^n)$ cuts (as opposed to $\Theta(4^n)$ cuts necessary for the description of $\widehat{\Lambda}^{\mathsf{CA}}(n)$ for the extended case).

6.5.2 Design Guidelines

The results presented in this chapter suggest the following design guidelines for utilizing caches in large wireless networks. First, as we have seen in Example 6.2, the strategy of selecting only the closest cache can be quite bad in general. Instead, the cache selection has to be done in a load-balanced fashion. In our proposed three-layer architecture, cache selection is performed by solving a linear program. Moreover, we argued that the solution to this linear program can be found efficiently by a distributed algorithm. On the other hand, when the cache location and traffic are already balanced enough (as is the case in Example 6.3), this load balancing can be omitted, and closest cache selection can indeed be close to optimal.

Second, the results suggest that, at least in the large path-loss regime, the problem of optimal cache selection can be solved at the routing layer. In other words, the physical and cooperation layer are not affected by the introduction of caches.

6.6 Chapter Summary

We analyzed the influence of caching on the performance of wireless networks. Our approach is information-theoretic, yielding a scaling characterization of the complete caching capacity region in the high path-loss regime $\alpha > 6$. Even though this region is $2^n \times n$ dimensional (i.e., exponential in the number of nodes n in the wireless network), we present an algorithm that checks approximate feasibility of a particular caching traffic matrix efficiently (in polynomial time in the description length of the caching traffic matrix). Achievability is proved using a three-layer communication architecture achieving the entire caching capacity region in the scaling sense for $\alpha > 6$. The three layers deal with optimal selection of caches, choice of amount of necessary cooperation, noise and interference, respectively. The matching (in the scaling sense) converse proves that addressing these questions separately is without loss of order-optimality in the regime of high path-loss exponent.

Chapter 7

Conclusions

This chapter contains concluding remarks. In Section 7.1 we summarize the results presented in this thesis. Section 7.2 contains pointers for future work.

7.1 Thesis Summary

In this thesis, we have considered the impact of heterogeneities on achievable rates in large wireless networks. Three types of such heterogeneities were discusses in detail: location heterogeneity, traffic heterogeneity, and service heterogeneity.

We analyzed location heterogeneity be allowing the nodes in the wireless network to be placed arbitrarily (with a minimum-separation constraint). This contrasts with the standard homogeneity assumption, i.e., that nodes are placed independently and uniformly at random. For the traffic model, we assumed random source-destination pairing with uniform rate. We have seen that the impact of arbitrary node placement depends strongly on the path-loss exponent α. For small path-loss exponents $\alpha \in (2, 3]$, we showed that the node placement has no impact on achievable rates and that global cooperation is necessary irrespective of the node placement. We proposed a novel cooperative communication scheme (called hierarchical relaying), and showed that it is order optimal for all node placements. For large path-loss exponents $\alpha > 3$, we have seen that the node placement critically impacts achievable rates as well as optimal communication schemes. For very regular node placements, multi-hop com-

munication is order optimal, for very irregular node placements, hierarchical relaying is optimal. We proposed a communication scheme "interpolating" between these two schemes depending on the regularity of the node placement, and showed that this scheme is order optimal under adversarial node placement with regularity constraint.

For traffic heterogeneity, we analyzed the n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$. In other words, we allowed general unicast traffic. This contrasts with the standard homogeneity assumption, i.e., that each node is source exactly once for a destination chosen independently and uniformly at random from among all the other nodes, and that all these n source-destination pairs communicate at equal rate. For the node placement, we assumed that nodes are place independently and uniformly at random. We presented inner an outer bounds on the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$. These bounds coincide up to a factor $n^{\pm o(1)}$ along at least $n^2 - n$ out of n^2 dimensions (corresponding to balanced traffic) for $\alpha \in (2, 5]$, and along all n^2 dimensions for $\alpha > 5$. Hence, for $\alpha > 5$, this provides a scaling characterization for the entire n^2-dimensional unicast capacity region $\Lambda^{\mathsf{UC}}(n)$, and the same statement is true along at least $n^2 - n$ dimensions for $\alpha \in (2, 5]$. For the inner bound, we provided a three-layer communication architecture. The three layers are the routing layer, the cooperation layer, and the physical layer. In the routing layer, we perform load balancing — dealing with the traffic heterogeneity. In the cooperation layer, we distribute and concentrate traffic over the wireless network — choosing the appropriate amount of cooperation. In the physical layer, we implement this distribution an concentration of traffic — handling interference and noise. The approximate optimality of this scheme (in the sense mentioned above) shows that these problems can be solved separately without loss of order optimality.

For service heterogeneity, we analyzed the $n \times 2^n$-dimensional multicast capacity region $\Lambda^{\mathsf{MC}}(n)$ and the $2^n \times n$-dimensional caching capacity region $\Lambda^{\mathsf{CA}}(n)$. This contrasts with the standard service homogeneity assumption, i.e., that all demands are unicast. For the node placement, we again assumed that nodes a placed independently and uniformly at random. We presented inner and outer bounds on $\Lambda^{\mathsf{MC}}(n)$ and $\Lambda^{\mathsf{CA}}(n)$. For the multicast case, these bounds coincide up to a factor $n^{\pm o(1)}$ along

at least $n2^n - n$ out of $n2^n$ dimensions for $\alpha \in (2, 5]$, and along all $n2^n$ dimensions for $\alpha > 5$. Hence, for $\alpha > 5$, this provides a scaling characterization of the entire $n \times 2^n$-dimensional multicast capacity region $\Lambda^{\mathsf{MC}}(n)$, and the same statement is true along at least $n2^n - n$ dimensions for $\alpha \in (2, 5]$. For the caching case, these bounds coincide up to a factor $n^{\pm o(1)}$ for $\alpha > 6$, yielding a scaling characterization of the entire $2^n \times n$-dimensional caching capacity region $\Lambda^{\mathsf{CA}}(n)$ in the high path-loss regime. For the inner bounds, we showed how the three-layer architecture proposed for general unicast traffic can be modified to accommodate multicast and caching traffic. In both cases, only the routing layer needed to be changed — the cooperation and physical layer operate as in the unicast case.

7.2 Future Work

There are several directions for future work, of theoretical as well as practical interest.

The following questions are of theoretical interest, aiming at broadening our understanding of large heterogeneous wireless networks. First, note that the converse for location heterogeneity in the $\alpha > 3$ regime is only under *adversarial* node placement with regularity constraint. One direction would be to find a converse for *any* node placement. Second, for both unicast and multicast traffic, the scaling characterization holds for all but n dimensions of the respective capacity regions in the $\alpha \in (2, 5]$ regime. A second direction for future work is to complete the scaling characterization for the remaining n dimensions in the low path-loss regime. Similarly, for caching traffic, the results are only partial (namely only an inner bound is available) for $\alpha \leq 6$. Completing the picture for $\alpha \leq 6$ would again be of interest. Third, the results for traffic and service heterogeneity are derived assuming random node placement. Obtaining scaling characterizations of the unicast or multicast capacity regions under arbitrary node placement would be of interest. Ultimately, the goal is to develop a complete scaling description of achievable rates in large heterogeneous wireless networks.

The following questions and directions for future work are of practical interest.

First, most of the architectures presented in this thesis are centralized. To be implementable in practice, decentralized architectures need to be developed. Second, issues of delay have been completely ignored throughout this thesis. Especially in the broadcast phase of the hierarchical relaying scheme, delays can be quite large. This issue needs to be addressed before being implementable. Third, how to implement the communication architectures proposed in this thesis within existing protocol stacks needs to be addressed. For example, the distribution and concentration of traffic in the cooperation layer of the three-layer architecture for traffic heterogeneity results in growing header overhead. The design guidelines presented towards the end of the various chapters may be helpful in deciding which parts of the communication schemes presented in this thesis are crucial for optimal operation of large wireless networks, and which parts can be modified for simpler implementation.

Bibliography

[1] Shuchin Aeron and Venkatesh Saligrama. Wireless ad hoc networks: Strategies and scaling laws for the fixed SNR regime. *IEEE Transactions on Information Theory*, 53(6):2044–2059, June 2007.

[2] Amit Agarwal, Noga Alon, and Moses S. Charikar. Improved approximation for directed cut problems. In *Proceedings of the ACM Symposium on Theory of Computing*, pages 671–680, 2007.

[3] Ashish Agarwal and P. R. Kumar. Capacity bounds for ad hoc and hybrid wireless networks. *ACM SIGCOMM Computer Communications Review*, 34(3):71–81, July 2004.

[4] Sahand Haji Ali Ahmad, Aleksandar Jovičić, and Pramod Viswanath. On outer bounds to the capacity region of wireless networks. *IEEE Transactions on Information Theory*, 52(6):2770–2776, June 2006.

[5] Joon Ahn and Bhaskar Krishnamachari. Fundamental scaling laws for energy-efficient storage and querying in wireless sensor networks. In *Proceedings of the ACM MobiHoc*, pages 334–343, May 2006.

[6] Ivan Baev, Rajmohan Rajaraman, and Chaitanya Swamy. Approximation algorithms for data placement problems. 38(4):1411–1429, August 2008.

[7] João Barros and Sergio D. Servetto. Network information flow with correlated sources. *IEEE Transactions on Information Theory*, 52(1):155–170, January 2006.

[8] Thomas M. Cover, Abbas El Gamal, and Masoud Salehi. Multiple access channels with arbitrarily correlated sources. *IEEE Transactions on Information Theory*, 26(6):648– 657, November 1980.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[10] Massimo Franceschetti, Olivier Dousse, David N. C. Tse, and Patrick Thiran. Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Transactions on Information Theory*, 53(3):1009–1018, March 2007.

[11] Massimo Franceschetti, Marco D. Migliore, and Paolo Minero. The capacity of wireless networks: Information-theoretic and physical limits. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, September 2007.

[12] Massimo Franceschetti, Marco D. Migliore, and Paolo Minero. The degrees of freedom of wireless networks: Information theoretic and physical limits. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, September 2008.

[13] Naveen Garg, Vijay V. Vazirani, and Mihalis Yannakakis. *Primal-dual approximation algorithms for integral flow and multicut in trees, with applications to matching and set cover*, pages 64–75. Lecture Notes in Computer Science. Springer, 1993.

[14] Piyush Gupta. *Design and Performance Analysis of Wireless Networks*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[15] Piyush Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, March 2000.

[16] Piyush Gupta and P. R. Kumar. Towards and information theory of large networks: An achievable rate region. *IEEE Transactions on Information Theory*, 49(8):1877–1894, August 2003.

[17] Te Sun Han. Slepian-Wolf-Cover theorem for network of channels. *Information and Control*, 47(1):67–83, January 1980.

[18] Shunsuke Ihara. On the capacity of channels with additive non-Gaussian noise. *Information and Control*, 37(1):34–39, April 1978.

[19] Sang-Woon Jeon and Sae-Young Chung. Two-phase opportunistic broadcasting in large wireless networks. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 2771–2775, June 2007.

[20] Shudong Jin and Limin Wang. Content and service replication strategies in multi-hop wireless mesh networks. pages 79–86, October 2005.

[21] Aleksandar Jovičić, Pramod Viswanath, and Sanjeev R. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Transactions on Information Theory*, 50(11):2555–2565, November 2004.

[22] Nabil Kahale. On reducing the cut ratio to the multicut problem. *Unpublished Manuscript*, 1993.

[23] Alireza Keshavarz-Haddad, Vinay Ribeiro, and Rudolf Riedi. Broadcast capacity in multihop wireless networks. In *Proceedings of the ACM MobiCom*, pages 239–250, September 2006.

[24] Alireza Keshavarz-Haddad and Rudolf Riedi. On the broadcast capacity of multihop wireless networks: Interplay of power, density and interference. In *Proceedings of the IEEE SECON*, pages 324–323, June 2007.

[25] Alireza Keshavarz-Haddad and Rudolf Riedi. Multicast capacity of large homogeneous multihop wireless networks. In *Proceedings of the IEEE WiOpt*, April 2008.

[26] Ashish Khisti, Uri Erez, and Gregory W. Wornell. Fundamental limits and scaling behavior of cooperative multicasting in wireless networks. *IEEE Transactions on Information Theory*, 52(6):2762–2769, June 2006.

[27] Bong-Jun Ko and Dan Rubenstein. Distributed self-stabilizing placement of replicated resources in emerging networks. *IEEE/ACM Transactions on Networking*, 13(3):476–487, June 2005.

[28] Gerhard Kramer, Michael Gastpar, and Piyush Gupta. Cooperative strategies and capacity theorems for relay networks. *IEEE Transactions on Information Theory*, 51(9):3037–3063, September 2005.

[29] Christof Krick, Harald Räcke, and Matthias Westermann. Approximation algorithms for data management in networks. 36(5):497–519, October 2003.

[30] Sanjeev R. Kulkarni and Pramod Viswanath. A deterministic approach to throughput scaling in wireless networks. *IEEE Transactions on Information Theory*, 50(6):1041–1049, June 2004.

[31] Tom Leighton and Satish Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pages 422–431, October 1988.

[32] Olivier Lévêque and İ. Emre Telatar. Information-theoretic upper bounds on the capacity of large extended ad hoc wireless networks. *IEEE Transactions on Information Theory*, 51(3):858–865, March 2005.

[33] Xiang-Yang Li, Shao-Jie Tang, and Ophir Frieder. Multicast capacity for large scale wireless ad hoc networks. In *Proceedings of the ACM MobiCom*, pages 266–277, September 2007.

[34] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, June 1995.

[35] Ritesh Madan, Devavrat Shah, and Olivier Lévêque. Product multicommodity flow in wireless networks. *IEEE Transactions on Information Theory*, 54(4):1460–1476, April 2008.

[36] Pavan Nuggehalli, Vikram Srinivasan, and Carla-Fabiana Chiasserini. Energy-efficient caching strategies in ad hoc wireless networks. In *Proceedings of the ACM MobiHoc*, pages 25–34, June 2003.

[37] Ayfer Özgür, Olivier Lévêque, and Emmanuel Preissmann. Scaling laws for one- and two-dimensional random wireless networks in the low-attenuation regime. *IEEE Transactions on Information Theory*, 53(10):3573–3585, October 2007.

[38] Ayfer Özgür, Olivier Lévêque, and David N. C. Tse. Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Transactions on Information Theory*, 53(10):3549–3572, October 2007.

[39] Srinivas Shakkottai, Xin Liu, and R. Srikant. The multicast capacity of large multihop wireless networks. In *Proceedings of the ACM MobiCom*, pages 247–255, September 2007.

[40] Birsen Sirkeci-Mergen and Michael Gastpar. On the broadcast capacity of wireless networks. In *Proceedings of the Information Theory and Applications Workshop*, January 2007.

[41] R. Srikant. *The Mathematics of Internet Congestion Control*. Birkhäuser, 2003.

[42] Sundar Subramanian, Sanjay Shakkottai, and Piyush Gupta. On optimal geographic routing in wireless networks with holes and non-uniform traffic. In *Proceedings of the IEEE INFOCOM*, May 2007.

[43] Sundar Subramanian, Sanjay Shakkottai, and Piyush Gupta. Optimal geographic routing for wireless networks with near-arbitrary holes and traffic. In *Proceedings of the IEEE INFOCOM*, pages 2002–2010, April 2008.

[44] Bin Tang, Samir Das, and Himanshu Gupta. Cache placement in sensor networks under update cost constraint. In *Ad-Hoc, Mobile, and Wireless Networks*, pages 334–348. Springer, 2005.

[45] Bin Tang, Himanshu Gupta, and Samir R. Das. Benefit-based data caching in ad hoc networks. *IEEE Transactions on Mobile Computing*, 7(3):289–304, March 2008.

[46] Bulent Tavli. Broadcast capacity of wireless networks. *IEEE Communications Letters*, 10(2):68–69, February 2006.

[47] L. G. Valiant and G. J. Brebner. Universal schemes for parallel communication. In *Proceedings of the ACM Symposium on Theory of Computing*, pages 263–277, 1981.

[48] Liang-Liang Xie and P. R. Kumar. A network information theory for wireless communication: Scaling laws and optimal operation. *IEEE Transactions on Information Theory*, 50(5):748–767, May 2004.

[49] Liang-Liang Xie and P. R. Kumar. An achievable rate for the multiple-level relay channel. *IEEE Transactions on Information Theory*, 51(4):1348–1358, April 2005.

[50] Liang-Liang Xie and P. R. Kumar. On the path-loss attenuation regime for positive cost and linear scaling of the transport capacity in wireless networks. *IEEE Transactions on Information Theory*, 52(6):2313–2328, June 2006.

[51] Feng Xue, Liang-Liang Xie, and P. R. Kumar. The transport capacity of wireless networks over fading channels. *IEEE Transactions on Information Theory*, 51(3):834–847, March 2005.

[52] Rong Zheng. Information dissemination in power-constrained wireless networks. In *Proceedings of the IEEE INFOCOM*, pages 1–10, April 2006.