Foundations and Trends® in
Networking
Vol. 5, No. 1 (2010) 1–107

DOI: 10.1561/1300000016

now
the essence of knowledge

Operating Regimes of Large Wireless Networks

By Ayfer Özgür, Olivier Lévêque and David Tse

Contents

Foundations and Trends® in
Networking
Vol. 5, No. 1 (2010) 1–107

DOI: 10.1561/1300000016

now
the essence of knowledge

Operating Regimes of Large Wireless Networks

Ayfer Özgür[1], Olivier Lévêque[2] and David Tse[1]

[1] *Ecole Polytechnique Fédérale de Lausanne, Faculté Informatique et Communications, Lausanne, 1015, Switzerland, {ayfer.ozgur,olivier.leveque}@epfl.ch*
[2] *University of California at Berkeley, Department of EECS, Berkeley, CA, 94720, USA, dtse@eecs.berkeley.edu*

Abstract

Multi-hop is the current communication architecture of wireless mesh and *ad hoc* networks. Information is relayed from each source to its destination in successive transmissions between intermediate nodes. A major problem regarding this architecture is its poor performance at large system size: as the number of users in a wireless network increases, the communication rate for each user rapidly decreases. Can we design new communication architectures that significantly increase the capacity of large wireless networks?

In this monograph, we present a scaling law characterization of the information-theoretic capacity of wireless networks, which sheds some light on this question. We show that the answer depends on the parameter range in which a particular network lies, namely the operating regime of the network. There are operating regimes where the information-theoretic capacity of the network is drastically higher than the capacity of conventional multi-hop. New architectures can provide

substantial capacity gains here. We determine what these regimes are and investigate the new architectures that are able to approach the information-theoretic capacity of the network. In some regimes, there is no way to outperform multi-hop. In other words, the conventional multi-hop architecture indeed achieves the information-theoretic capacity of the network. We discuss the fundamental factors limiting the capacity of the network in these regimes and provide an understanding of why conventional multi-hop indeed turns out to be the right architecture.

The monograph is structured as follows: In Section 2, we discuss the role of interference in wireless networks. We show that while current communication architectures are fundamentally limited by interference, new architectures based on distributed MIMO communication can overcome this interference limitation, yielding drastic performance improvements. Section 3 discusses the impact of power. We show that in power-limited regimes, distributed MIMO-based techniques are important not only because they remove interference but also because they provide received power gain. We identify the power-limited operating regimes of wireless networks and define the engineering quantities that determine the operating regime of a given wireless network. We show that unless the wireless network operates in a severely power-limited regime, distributed MIMO communication provides significant capacity gain over current techniques. Finally, in Section 4, we study how the area of the network, i.e., space, impacts the capacity of the network. This study enriches the earlier picture by adding new operating regimes where wireless networks can be moderately or severely space-limited. We see that unless the network is severely limited in space, distributed-MIMO-based communication continues to provide drastic improvements over conventional multi-hop.

1

Introduction

In wired networks, a source can send information to a destination by routing it along a path, where intermediate nodes forward the information towards the destination. The application of this strategy to wireless networks has been the subject of a large body of research in the past two decades. Similar to wired networks, packets are sent here from each source to its destination via multiple intermediate nodes acting as relays. Each relay decodes the packets sent from the previous relay and forwards them to the next.

Multi-hop is a natural fit for wired networks; however, it is not clear whether it provides a good premise for wireless. It is based on point-to-point communication between nodes. Wired networks are already composed of point-to-point links over which signals travel in isolation. However, the notion of a point-to-point link is vague in the case of wireless.

Wireless signals are not isolated and they interact in complex ways. The signal transmitted by a given user is heard not only by its intended receiver but also by all the receivers in the vicinity of the transmitter. When there are multiple simultaneous transmissions over the same frequency band, each receiver observes a mixture of all the transmitted

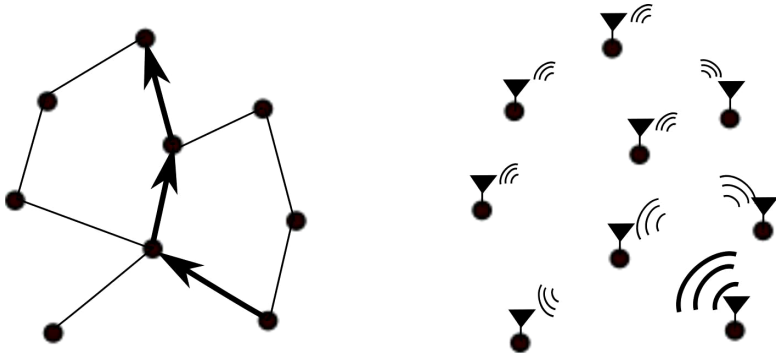

Fig. 1.1 Wired vs. wireless networks.

signals. Therefore, signals of interest to a receiver mix together with overheard signals from other transmissions. As wired and wireless networks are so different in their fundamental nature, it is not clear whether an architecture rooted in the practice of the former can provide a good premise for the latter. (Figure 1.1).

Today, there is an increasing need to connect a massive number of wireless devices and to support various resource-intensive applications. This leads us to especially discuss the performance of the conventional multi-hop architecture in large wireless networks: Can multi-hop efficiently support communication in large wireless networks or do we need new architectures for the rapidly growing wireless networks of the future? In particular, can new architectures tailored for wireless significantly outperform multi-hop in large networks? In this monograph, we study the information-theoretic capacity of large wireless networks, to shed some light on these questions.

1.1 Interference

In this section, we argue that the performance of the conventional multi-hop architecture is fundamentally limited by the interference between simultaneous transmissions in the shared wireless medium. However, this interference limitation is not fundamental and can be overcome with new architectures tailored for wireless. We discuss hierarchical cooperation, an architecture that constructively uses

interference for communication. As a result, it offers significant performance gains in large networks. This section provides a summary of Section 2 of this monograph.

1.1.1 Multi-hop is Interference Limited

Multi-hop is based on relaying information from sources to destinations via successive point-to-point transmissions between intermediate nodes. To do the point-to-point transmissions, we need to designate nodes in the network as transmitter–receiver pairs. Each receiver is to decode the message from its designated transmitter. Overheard signals from other transmitters constitute harmful *interference* corrupting the desired signals and are treated as additional noise at the receivers. The choice of these transmitter–receiver pairs in the network is a major optimization problem determining the throughput performance of the multi-hop architecture.

In order to achieve high overall throughput, it is desired to choose the transmitter–receiver pairs such that many of them can communicate simultaneously without interfering too much with each other. This would provide a dense mesh for relaying information inside the network. On the other hand, it is also desirable to have a large separation between every transmitter–receiver pair so that messages advance by a large distance towards their destinations in every hop. The interference between simultaneous transmissions poses a fundamental trade-off between these two trends. Each transmission creates strong interference for other receivers around its transmitter. The radius of this strong interference zone is proportional to the transmitted power, which is in turn proportional to the range of the targeted transmission (Figure 1.2). Therefore, the larger the separation between the transmitter-receiver pairs in the network, the fewer of them can communicate at the same time.

In particular, if we allow for direct transmissions from the source nodes in the network to their destinations, only few of these source-destination pairs can communicate at a time, as source-destination pairs in a network are typically separated by large distances. Consider a network with a large number of users n, where users are randomly

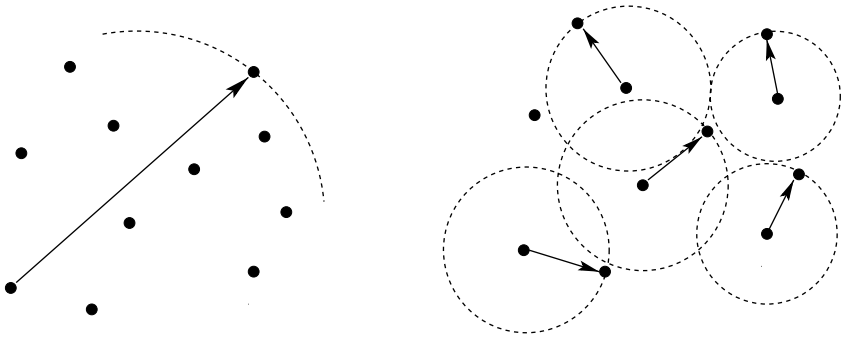

Fig. 1.2 Long vs. short-range communication in wireless networks. The nodes inside each circle are subject to interference from the corresponding transmission.

paired into $n/2$ source-destination pairs. Each source wants to communicate to its corresponding destination node. Such a random pairing will lead to $\Theta(n)$ pairs separated by a distance of the order of the diameter of the network. If source-destination pairs are to communicate directly with each other, these $\Theta(n)$ pairs should go one at a time. The per-pair communication rate with such a time-sharing strategy decreases as $\Theta(1/n)$ with increasing number of users n. Note that each pair gets to transmit once in $\Theta(n)$ time slots.[1]

The other extreme is to confine to nearest-neighbor communication inside the network. As wireless signals get attenuated with distance, many local communications can be simultaneously active without interfering too much with each other (spatial reuse). See Figure 1.2. In particular, confining to nearest-neighbor communication maximizes the number of simultaneous transmissions inside the network. However, to cover long distances in short hops, each packet now has to be retransmitted many times before getting to its final destination. This relaying burden limits the achievable throughput.

In their seminal work [9] in 2000, Gupta and Kumar showed that confining to nearest-neighbor transmissions maximizes the throughput of multi-hop and provides an aggregate throughput of order $\Theta(\sqrt{n})$ in a network of n users. This corresponds to a per-user rate that

[1] Here, transmissions are orthogonalized over time so that they do not interfere. Equivalently, transmissions can be orthogonalized in frequency or in code space.

scales as $\Theta(1/\sqrt{n})$. Note that this scaling is significantly better than the $\Theta(1/n)$ scaling with direct communication (single-hop) between source-destination pairs. Nevertheless, it still decreases quite rapidly to zero with an increasing number of users n. This limitation is precisely due to the fact that in a nearest-neighbor multi-hop architecture, most users have to relay information for $\Theta(\sqrt{n})$ source-destination pairs on average.

The sub-linear scaling of the system throughput is fundamentally due to the need to reduce interference between point-to-point transmissions. If transmissions were not interfering, we could have many simultaneous long distance transmissions in the network, ideally every source could directly and simultaneously communicate to its destination. It is because of the interference that we need to confine to short distance communication, in which case the resulting relaying burden limits the system throughput.

1.1.2 Constructive Use of Interference: Hierarchical Cooperation

A natural question is whether we can surpass the interference barrier by allowing more sophisticated cooperation between the nodes, in particular by removing the restriction to point-to-point communication. Can we design cooperation architectures whose performance *scales* with system size? In Section 2, we present a hierarchical cooperation architecture that achieves an aggregate throughput of $\Theta(n^{1-\epsilon})$ for any $\epsilon > 0$. An aggregate throughput scaling arbitrarily close to linear in the number of nodes means that there is essentially no interference limitation: The rate for each source-destination pair does not degrade significantly, even if the network serves a growing number of users. This result demonstrates that the fundamental capacity of wireless networks can be significantly higher than the capacity of multi-hop and that more sophisticated cooperation architectures can provide substantial performance gains in large networks.

The key to this result is distributed MIMO (multi-input multi-output) communication. MIMO is a physical-layer technique, which was originally developed in the classical point-to-point setting. In this

setting, multiple antennas are installed on both the transmitter and the receiver. This allows to simultaneously send an independent stream of data from each transmit antenna. Each receive antenna observes a different combination of the transmitted signals. Jointly processing the vector of received observations at the antennas allows the receiver to remove the interference between the transmitted signals and recover the original data streams [5, 29]. A natural approach to apply this concept to the network setting is to have nodes cooperate in *clusters* to form distributed transmit and receive antenna arrays. In this manner, mutually interfering signals can be turned into useful ones that can be jointly decoded at the receive cluster and spatial multiplexing gain can be realized.

One way to incorporate distributed MIMO communication is to transfer the packets of each source node to its destination in three consecutive phases: The packets of a source node are first distributed among a cluster of nodes in its vicinity. In a second phase, the nodes in this source cluster simultaneously transmit these packets to a group of nodes around the destination node. These simultaneous transmissions can be regarded as distributed MIMO communication if the observations of the various nodes in the destination cluster can be jointly processed. Therefore, in a third phase, the distributed MIMO observations should be collected at the actual destination node, which can then jointly process these observations and recover the packets from its source node.

The above strategy potentially offers performance gain via the simultaneous long distance transmissions in the second phase. The interference between these transmissions is not anymore harmful, as they are jointly decoded at the end of the third phase. However, the overhead introduced by the first and the third phases to establish the necessary transmit and receive cooperation can drastically reduce the useful throughput. The key to efficient cooperation in the first and third phases is a digital and *hierarchical* architecture that makes use of distributed MIMO communication at increasing scales. Cooperation first takes place between nodes within small local clusters. These small clusters can operate simultaneously, as the decay of signals with distance allows spatial reuse. The cooperation facilitates MIMO

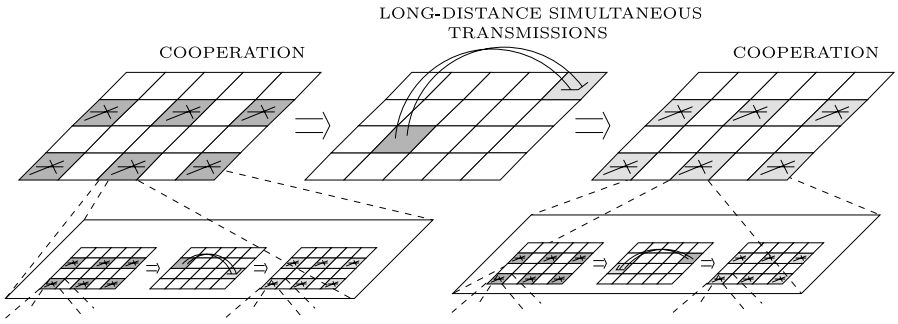

Fig. 1.3 The salient features of the hierarchical cooperation architecture.

communication over a larger spatial scale. This can then be used as a communication infrastructure for cooperation within larger clusters at the next level of the hierarchy. Continuing in this fashion, cooperation can be achieved at an almost global scale. At the highest level of the hierarchy, long-range MIMO communications can be performed between clusters almost as large as the whole network. By increasing the number of levels in such a hierarchical architecture, one can get arbitrarily close to linear aggregate throughput scaling. Figure 1.3 illustrates the hierarchical architecture with a focus on the top two levels.

The distributed MIMO-based approach summarized above is closely related to physical-layer network coding. Physical-layer network coding [11, 34] is another recent paradigm in wireless networking, based on the same motivation to embrace the wireless interference instead of avoiding it. Physical-layer network coding allows for two strategically picked transmissions to interfere at a relay node, which then forwards the mixture of the two signals. The fundamental difference between distributed MIMO-based hierarchical cooperation and physical-layer network coding is the scale over which wireless interference is embraced. Physical-layer network coding maintains the multi-hop architecture at the global scale and allows two local transmissions to interfere at each hop. Such an approach has the potential to double the throughput of the network, but no more (this was shown precisely in Ref. [17]). In the hierarchical cooperation architecture, communication is organized so that wireless interference can be embraced at the global scale.

It can be viewed as an aggressive form of physical-layer network coding, where $\Theta(n)$ transmissions are allowed to interfere instead of only two. Consequently, the gain is more substantial: instead of doubling the aggregate throughput, we can elevate its scaling from $\Theta(\sqrt{n})$ to linear in n.

1.2 Power

Interference is not the only factor that can potentially limit performance in wireless networks. Power can be another limiting factor. In some wireless networks, the reason to confine to short-range communication and relay packets via multiple transmissions may not be the interference that would be caused by long distance communication. The attenuation of wireless signals with distance may not allow sufficient received SNR (signal-to-noise power ratio) to directly reach far-away destinations. This can be the case due to a number of reasons:

(a) The power available at the nodes can be limited.
(b) The network can be distributed over a large geographical area.
(c) The attenuation in the environment can be high.
(d) The network can be operating on a large bandwidth (wideband system).

The objective in such wireless networks is not only to deal with interference but also to transfer power efficiently to the receivers. In particular, in an extremely power-limited network interference may be far below the noise level at the receivers. In such a regime, the strategies that provide the best throughput would be the ones that utilize power most efficiently. In this section, we discuss the question we raised in the previous section by also putting *power* into play: Is the traditional multi-hop architecture able to efficiently transfer power in large wireless networks? Can more sophisticated architectures, for example, hierarchical cooperation, provide significant capacity improvement in *power-limited* wireless networks?

The restriction to point-to-point communication in the traditional multi-hop setting can now be questioned from the *power* point of view.

In point-to-point communication, the signals received from a particular transmission are treated as noise at all but one receiver inside the network. In the previous section, we have seen that with distributed MIMO-based communication, we are able to turn mutually interfering signals into useful ones. By exploiting the broadcast nature of the wireless medium, such techniques can provide a received power gain, in addition to the spatial multiplexing gain emphasized in the previous section. This power gain can translate into a significant capacity gain in certain power-limited networks. The impact of power is discussed in detail in Section 3. We provide below a short summary of the conclusions of this section.

1.2.1 Impact of Power in the Point-to-point Wireless Channel

To understand the impact of power in wireless networks, let us first review how the amount of available transmission power impacts the capacity of the point-to-point additive white Gaussian noise channel. The capacity of this channel is given by Shannon's famous formula

$$C = W \log\left(1 + \frac{P}{N_0 W}\right) \tag{1.1}$$

in terms of the bandwidth of the channel W in Hz, received power P in Watts, and noise power spectral density $N_0/2$ in Watts/Hz.

The most important engineering parameter we associate with this channel is SNR defined as

$$\text{SNR} = \frac{P}{N_0 W}.$$

This parameter determines the operating regime of the channel. When $\text{SNR} \ll 0\,\text{dB}$, the channel is in a power-limited regime: the capacity is approximately linear in the power, and the performance depends critically on the power available, but not so much on the bandwidth. In this regime, if we double the transmit power, we can approximately double the channel capacity; however, doubling the bandwidth only marginally improves capacity. In the bandwidth-limited (or high-SNR) regime, where $\text{SNR} \gg 0\,\text{dB}$, we have the opposite situation: the capacity is

approximately linear in the bandwidth and the performance depends critically on the bandwidth, but not so much on the power. These two observations can be immediately verified from the capacity formula in Equation (1.1), noting that when $\text{SNR} \ll 0\,\text{dB}$, $\log(1+x) \approx x$ and when $\text{SNR} \gg 0\,\text{dB}$, the logarithm function gets saturated and increases very slowly in its argument.

These two fundamentally different operating regimes have two completely different implications in terms of communication system design. For a bandwidth-limited channel, the least we would expect from a good communication strategy for this channel is that its performance is approximately linear in the bandwidth, i.e., able to follow the trend of the capacity. On the contrary, for a power-limited channel, we should design a strategy whose performance increases linearly in the power. In the sequel, we will call a strategy scaling optimal or simply optimal for a certain regime, if its performance exhibits approximately the same dependence to system parameters as the information-theoretic capacity of the system.[2] Note that there is no guarantee that a strategy which is scaling optimal for a certain regime, meaning that its performance exhibits approximately the right behavior in terms of system parameters in this regime, would also be optimal for another regime.

1.2.2 Impact of Power in Wireless Networks

The interference discussion of the earlier section was implicitly based on a regime where the capacity of the wireless network is bandwidth-limited. The basis for the discussion was the scaling law approach of Gupta and Kumar [9], which looks at how the capacity of the network scales with the number of users. As the number of users in the wireless network increases, the other parameters of the network, such as area, bandwidth, per-user power, are kept fixed. This scaling results in a large network whose information-theoretic capacity is approximately given by nW. While the capacity of multi-hop in this regime behaves as $\sqrt{n}W$, the capacity of the new hierarchical cooperation strategy behaves as nW. This makes hierarchical cooperation scaling optimal in this regime.

[2] The approximation is within a poly-logarithmic factor.

The discussion on the operating regimes of the point-to-point wireless channel suggests that we could also have power-limited operating regimes in wireless networks. In this case, however, the capacity exhibits a completely different behavior. Indeed, power turns out to be a more sophisticated player in wireless networks than in the point-to-point case. There are a number of fundamentally different power-limited regimes in wireless networks. This is first due to the fact that the power limitation is jointly determined by a number of independent parameters (a)–(d) listed above. These parameters have different impact and their interplay creates a number of qualitatively different cases. For example, a network that suffers power limitation due to high attenuation in the environment is not equivalent to (cannot be translated to) a network that suffers from limited power available at the wireless nodes. Second, a wireless network can be power-limited in different degrees. For example, in a severely power-limited wireless network, channels between all pairs of nodes in the network are weak (of low SNR). In less severe cases, only the channels between far-away pairs are weak, whereas close-by nodes are connected via strong channels (of high SNR).

The backbone of the hierarchical cooperation architecture introduced in the previous section is distributed MIMO communication: at the highest level of the hierarchy, we perform simultaneous long distance transmissions from a source cluster of $\Theta(n)$ nodes to a destination cluster of $\Theta(n)$ nodes. The transmissions from each node in the source cluster are heard by all the nodes inside the destination cluster, though these $\Theta(n)$ simultaneous transmissions interfere with each other. When the interference between these transmissions is removed via joint decoding at the destination node, power-wise, it is as if we were able to observe each transmission interference-free at $\Theta(n)$ different receivers. In other words, for each transmission, the hierarchical cooperation architecture collects the power received by the $\Theta(n)$ nodes inside the destination cluster.

This leads to the following interesting fact: *A priori*, we may expect to observe some sort of power limitation in a wireless network if the received SNR between some pair of nodes in the network is not sufficient for direct communication, most notably between far-away

pairs. However, the information-theoretic capacity of the network is bandwidth-limited and not power-limited, approximately given by nW, as long as n *times* the SNR between far-away pairs is larger than 0 dB. We define this quantity as the long distance SNR of the network, denoted as SNR_l: it is n *times* the received SNR of a point-of-point channel with the transmitter and receiver separated by a distance equal to the diameter of the network. Note that the diameter defines the largest geographical scale for communication inside the network. As long as $\mathrm{SNR}_l \gg 0$ dB, the wireless network is bandwidth-limited and hierarchical cooperation is scaling optimal. This is not only because hierarchical cooperation can handle interference efficiently as discussed in the earlier section but also because it is able to efficiently exploit the broadcasting nature of the wireless medium in this regime.

A wireless network starts to experience power limitation when the long distance SNR drops below 0 dB. When $\mathrm{SNR}_l \ll 0$ dB, the network is power-limited over the largest geographical scale but can still be bandwidth-limited over a shorter communication scale. The optimal cooperation architecture is determined by two parameters in this case. The first one is the power path loss exponent of the environment, α. It describes how fast signal power decays with distance: signals transmitted from one node to another at distance r apart are subject to a power loss of $r^{-\alpha}$, where typically $2 \leq \alpha \leq 6$. $\alpha = 2$ corresponds to free-space propagation and larger α to more lossy environment. The power path loss exponent defines a dichotomy: when $2 \leq \alpha < 3$, the hierarchical cooperation architecture transfers power optimally inside the network and achieves the information-theoretic capacity scaling of the network. Signal power decays slowly with distance in this case and hierarchical cooperation yields maximal received power by collecting the received signals of $\Theta(n)$ nodes around each destination node.

When $\alpha \geq 3$, signal power decays fast with distance and long distance communication in the network is not preferable, even with its additional $\Theta(n)$ power gain. The optimal architecture depends on the strength of the power limitation in the network, which is captured by a second SNR parameter, the short distance SNR, denoted by SNR_s. SNR_s is the received SNR in a point-to-point transmission over the typical nearest-neighbor distance inside the network. The nearest-neighbor distance is

the shortest scale for communication inside the network. When $\mathrm{SNR}_s \ll 0\,\mathrm{dB}$, communication over even the shortest geographical scale is limited in power. In this case, the conventional nearest-neighbor multi-hop architecture is the fundamentally right strategy for transferring power; it indeed achieves the information-theoretic scaling of the network capacity. The broadcasting nature of the wireless media plays insignificant role in such severely power-limited networks and therefore, confining to point-to-point communication is not anymore suboptimal.

When $\alpha \geq 3$, but $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$, the nearest-neighbor scale is bandwidth-limited. Note that $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$; hence, the network is still power-limited over distances of the order of the network diameter. In this case, the broadcasting of wireless signals is significant up to an intermediate geographical scale determined by the precise value of SNR_s and α. There is therefore the potential to improve performance with long distance communication up to this particular geographical scale. Beyond this scale, the network is power-limited and power attenuates rapidly for $\alpha > 3$; hence, communication over longer distances is inefficient. The optimal solution is to form MIMO clusters of an intermediate size and then multi-hop across several clusters to get to the final destination cluster. Each hop between adjacent clusters is now performed using distributed MIMO transmissions of the corresponding intermediate scale. This hybrid architecture is illustrated in Figure 1.4.

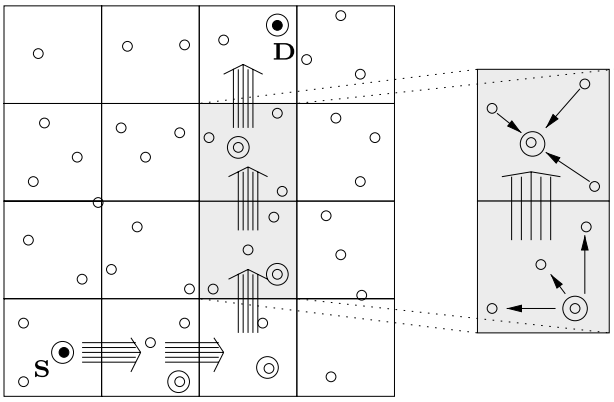

Fig. 1.4 Cooperate locally multi-hop globally: A generic optimal architecture for wireless networks.

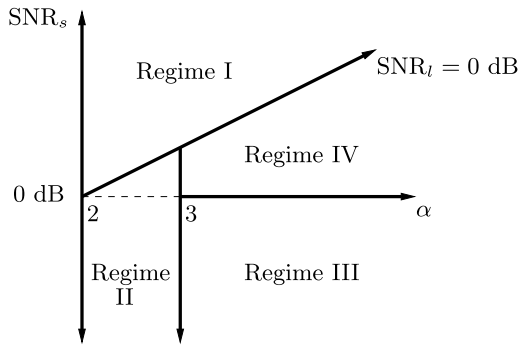

Fig. 1.5 The four operating regimes. The optimal schemes in these regimes are I–II: Hierarchical cooperation, III — Multi-hop, IV — Hybrid Multi-hop + Hierarchical Cooperation.

The two extremes of this architecture are precisely traditional multi-hop, where the cluster size is 1 and the number of hops is $\Theta(\sqrt{n})$, and hierarchical cooperation, where the cluster size is $\Theta(n)$ and the number of hops is 1. This hybrid architecture combining hierarchical cooperation with multi-hop provides a generic optimal solution for all wireless networks. For optimality, the cooperation scale should be adjusted according to the power available in the network and the power path loss exponent of the environment. The resultant four operating regimes and the corresponding optimal schemes for each regime are illustrated in Figure 1.5.

1.3 Space

The geographical area of the network not only plays a role in determining the received powers in the network, but also has an independent impact on capacity. It determines the number of independent spatial channels available for communication inside the wireless network. Information is communicated in the form of electromagnetic waves and the area of the network determines the diversity available in the physical channel. Consider the $\Theta(n)$ simultaneous long distance transmissions between the source and the destination clusters in the hierarchical cooperation architecture. Each node in the destination cluster observes a linear combination of the transmitted electromagnetic signals, each scaled and shifted according to the loss and delay in the corresponding

path. The destination node can only remove the interference between these transmissions via joint decoding, if the linear combinations of the signals are independent. When the $\Theta(n)$ nodes in the source cluster and the $\Theta(n)$ nodes in the destination cluster are packed together in small geographical areas, the linear combinations of the transmitted signals cannot be anymore independent. In this section, we reconsider the question raised in the earlier sections by concentrating on the impact of space on the capacity of wireless networks.

As we discuss in detail in Section 4, there are $\Theta(\sqrt{A}/\lambda)$ spatial degrees of freedom in a wireless network of area A, operating on a carrier wavelength λ. This is the number of independent spatial channels available for communication inside the network. Limited by interference, the multi-hop architecture can only achieve $\Theta(\sqrt{n})$ degrees of freedom. If the number of spatial degrees of freedom in the network is already as small as $\Theta(\sqrt{n})$, then multi-hop is fundamentally optimal, as it is able to achieve the full degrees of freedom of the network. When the available degrees of freedom in the network are more than $\Theta(\sqrt{n})$, there is the potential to exploit these additional degrees of freedom by more sophisticated cooperation. We will discuss in Section 4 that when the number of spatial degrees of freedom in the network is larger than $\Theta(\sqrt{n})$, the hierarchical cooperation architecture is able to achieve the full degrees of freedom in the network given by

$$\min(n, \sqrt{A}/\lambda).$$

In particular in wireless networks where $\sqrt{A}/\lambda \gg n$, there is no space limitation, as there are sufficient spatial degrees of freedom for all users. Hierarchical cooperation achieves linear aggregate throughput scaling in this case.

1.4 Operating Regimes

In this monograph, we discuss three factors that can potentially limit performance in wireless networks. We have already seen that the first one, interference, usually thought of to be a major performance limitation in wireless networks, can be overcome with cooperation between

nodes. The latter two, power and space, impose fundamental limitations on communication in wireless networks.

Can more sophisticated cooperation techniques provide significant capacity gains over the conventional multi-hop architecture in large wireless networks? We have seen that the answer to this question depends on the parameter range in which a particular wireless network lies. This naturally fits in a framework of operating regimes. Each operating regime corresponds to a subset of the parameter space where the optimal architecture for cooperation is different. The underlying reason is that the information-theoretic capacity of the network exhibits a qualitatively different behavior in each of these regimes. We have seen that there are many operating regimes where the information-theoretic capacity of the network is significantly higher than the capacity of conventional multi-hop and where architectures better tailored for wireless networks, hierarchical cooperation in particular, can provide substantial capacity gains. In certain regimes, most notably when the network is severely limited in either power or space, there is no way to outperform multi-hop. In other words, the conventional multi-hop architecture is able to achieve the information-theoretic scaling of the network capacity and is fundamentally optimal.

Which of these operating regimes are most relevant to practice? The above discussion also identifies the engineering quantities that determine the operating regime of a wireless network, such as short-range SNR, long-range SNR, area, power path loss exponent, etc. Note that these quantities can be easily computed or directly measured in the network. In Example 1.1 below, we plug in some typical values for the parameters of the network to get some insight on the most relevant operating regimes for various applications in practice.

Example 1.1. Suppose that, as a communication systems engineer, you need to suggest a communication architecture for a wireless network which will operate on a university campus. The campus has an area of $A = 1\,\text{km}^2$ and will operate around $3\,\text{GHz}$ ($\lambda = 0.1\,\text{m}$). According to the discussion in Section 1.3, the number of spatial degrees of freedom in the network is given by $\sqrt{A}/\lambda = 10'000$. Therefore, if there

are up to 10'000 students, we expect to have no space-limitation in the network: there are sufficient spatial degrees of freedom for all users to communicate. When there are more than 10'000 users, the network is space-limited. However, multi-hop can achieve all the degrees of freedom only when the number of users in this network of area $1\,\mathrm{km}^2$ are larger than 10^8, a humongous number. Up to this size, we need hierarchical cooperation to exploit the available degrees of freedom in the network. This suggests that although in practice, we might have wireless networks that are space-limited, severely space-limited networks where multi-hop is the right architecture are very unlikely.

In addition, we would most often expect such a network to be bandwidth-limited and not power-limited. Under free-space propagation, the transmitted power P and the received power P_r are related by the Friis formula:

$$P_r = \frac{G_{Tx} \cdot G_{Rx}}{(4\pi r/\lambda)^2} P,$$

where r is the distance between the transmitter and the receiver and G_{Tx} and G_{Rx} are the transmit and receive antenna gains. Assuming unit transmit and receive antenna gains, the attenuation factor $(G_{Tx} \cdot G_{Rx} \cdot \lambda^2)/16\pi^2$ in the formula is 10^{-6}. Assume transmitter power P of 100 mW per node, thermal noise N_0 at $-174\,\mathrm{dBm}$, a bandwidth W of 10 MHz and noise figure $\mathrm{NF} = 10\,\mathrm{dB}$. The SNR between a transmitter and receiver pair separated by the maximal distance of 1 km is 54 dB. With 10'000 users in the network, the long distance SNR is $\mathrm{SNR}_l = 104\,\mathrm{dB}$, very much in the high SNR regime. Note that even if the transmit power per node is 1 mW, a value more typical for sensor nodes, we still have $\mathrm{SNR}_l = 84\,\mathrm{dB}$, a bandwidth-limited network. In a lossy environment, SNR_l will be smaller, but can still be expected to be well above 0 dB.

Therefore, with 10'000 students on the campus, we do not expect to observe any power or space-limitation in the network. In this case, while traditional multi-hop can achieve a total throughput of the order of 100 bits/s/Hz, hierarchical cooperation promises an aggregate throughput of the order of 10'000 bits/s/Hz.

1.5 Problem Formulation

The results presented in this monograph are based on a scaling law characterization of the information-theoretic capacity of wireless networks. This scaling law formulation, developed mainly in Section 3.1, is used as a mathematical tool to identify the operating regimes of large wireless networks, without having to exactly characterize their capacity. It is based on identifying the parameters of wireless networks that have large operational range in practical applications, such as the area of the network, the transmit power available at the users and the bandwidth. Note that these are independent parameters, each of which can be large or small in different applications. As there are no typical values for these parameters, a thorough understanding of the capacity requires to study the interplay between these parameters. We model the interplay through a coupling to the number of users. Characterizing the scaling exponent of the capacity with the number of users for all possible couplings accounts for all possible interplay between these system parameters. Such a scaling law study allows not only to identify the operating regimes of wireless networks but also to approximately characterize the dependence of the information-theoretic capacity of the network to major system parameters.

There are two aspects to such a scaling exponent characterization: upper and lower bounds. Upper bounds on the best possible scaling exponent are derived using tools from information theory. Lower bounds are obtained by constructing explicit cooperation architectures and computing the scaling exponents they achieve. An architecture is called scaling optimal for a certain regime if it is able to achieve the best possible scaling exponent in this regime. This means that the performance of the architecture exhibits the same dependence to system parameters as the capacity itself. Such an optimality definition has an engineering significance: it guarantees that the gap to the information-theoretic capacity of the network does not explode rapidly with any of the system parameters.

The current text is slightly biased in detail towards lower bounds, as we believe the architectures themselves are of higher engineering interest than the theoretical proofs of their optimality. However, without

going into too much technical detail, we also tried to give the main intuition behind the information-theoretical upper bounds on capacity.

1.6 Historical Notes

The line of research that leads to the results summarized in this paper was initiated by the seminal work of Gupta and Kumar in 2000 [9]. The work of Gupta and Kumar was stimulating from several points of view. First, it initiated the study of the *scaling* of the capacity of wireless networks with the number of users. Such a scaling law formulation puts the emphasis on large system size and is useful to devise architectural guidelines for large wireless networks. The formulation turned out to be more amenable to analysis than the long-sought capacity region in information theory for a given number of users. Second, it introduced a simple random network model that captures the essential aspects of the problem: the spatial distribution of nodes over the network area and the traffic requirement between them, the attenuation of wireless signals with distance and the broadcasting and superposition nature of wireless media. Most importantly, using this model, Gupta and Kumar identified the interference-limited nature of the conventional multi-hop architecture, showing that in the best case, it achieves a $\Theta(\sqrt{n})$ scaling of the system throughput. A scheme achieving exactly $\Theta(\sqrt{n})$ throughput for generic random wireless networks was then proposed in Ref. [6].

The work of Gupta and Kumar inspired the research tackling the main question of interest in this monograph: Can we do better by more sophisticated physical-layer processing? This question was first addressed by Xie and Kumar [31]. They showed that whenever $\mathrm{SNR}_s \ll 0\,\mathrm{dB}$ and the power path loss exponent α of the environment is greater than 6, the nearest-neighbor multi-hop architecture is in fact order-optimal. The work [31] was followed by several others [1, 10, 16, 26, 32, 33]. Successively, they improved the threshold on the path loss exponent α for which multi-hop is order-optimal due to the severe power limitation $\mathrm{SNR}_s \ll 0\,\mathrm{dB}$ ($\alpha > 5$ in [10], $\alpha > 4.5$ in [1], $\alpha > 4$ in [32], and $\alpha > 3$ in [26]). The work of Franceschetti et al. [7] established the optimality of multi-hop under severe space-limitation (when $\sqrt{A}/\lambda \ll \sqrt{n}$). A similar conclusion was earlier obtained in [25]

by modeling the space limitation through a degenerate physical channel model.

Aeron and Saligrama were the first to show that the interference limitation suffered by conventional multi-hop can be surpassed with more sophisticated cooperation: they exhibited a scheme that yields a throughput scaling of $\Theta(n^{2/3})$. The hierarchical cooperation architecture achieving aggregate throughput $\Theta(n^{1-\epsilon})$ for any $\epsilon > 0$ has been introduced by the authors [26]. Both the scheme proposed by Aeron and Saligrama and the hierarchical cooperation architecture are based on combining MIMO communication [5, 29] with cooperative relaying ideas from network information theory. In particular, the hierarchical cooperation scheme critically employs compress-and-forward relaying, a strategy introduced in [4] (see also Refs. [12, 13]). The hybrid architecture combining hierarchical cooperation with multi-hop was introduced in Ref. [23]. The same paper also shows that this hybrid architecture surpasses multi-hop when the network is not severely power-limited, either when $\alpha < 3$ or when $\mathrm{SNR}_s \gg 0$ dB. The same hybrid architecture was independently proposed in Ref. [20] to deal with arbitrary placement of nodes inside the network area. The optimality of hierarchical cooperation when the network is partially space-limited was established in Refs. [14, 15, 27].

The characterization of wireless networks presented in this paper is based on the operating regimes framework developed in Ref. [23]. This framework offers a unified perspective on various fragmented or even seemingly contradicting results in this field. More importantly, it allows the deduction of concise engineering principles from the theory.

There are many interesting ideas we have not included in this monograph. In Refs. [20, 21], Niesen et al. extend some of the ideas in this monograph to networks with arbitrary node placement and arbitrary traffic demand. The work [24] investigates the throughput-delay tradeoff of the hierarchical cooperation scheme and [8] provides a refined analysis of its performance.

In an independent line of research, Cadambe and Jafar [3] and Nazer et al. [18] showed that interference alignment techniques provide an alternative way of dealing with interference in wireless networks and achieving high throughput. The scaling performance of these techniques

in wireless networks has been discussed in Refs. [19, 28]. The basic idea behind these techniques is fundamentally different from the schemes discussed in this monograph, in the sense that communication between order n source-destination pairs should be established in one shot, without intermediate relaying. Instead, by making use of sophisticated signaling at the transmitters, each destination receives a signal with two orthogonal components, one of which is the intended signal, whereas the other contains the interfering signals transmitted by all the other users. The intended signal can then recovered at each destination by a simple projection.

One of the major differences between interference alignment and cooperative schemes is therefore that interference alignment schemes do not rely on spatial reuse, which makes them superior in the regime when SNR is extremely large. On the other hand, they heavily rely on transmit channel state information, which is challenging to get in practice, while the techniques discussed in this monograph require channel state information only at the receiver side. Also, interference alignment techniques are less efficient in terms of power transfer than distributed MIMO transmissions: we have indeed seen above that distributed MIMO transmissions benefit from a significant power gain, of the order of the number of nodes participating to the transmission; this power gain is simply absent in interference alignment schemes.

1.7 Notation

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

2

Interference

In this section, we study the impact of interference on the capacity of wireless networks. We first highlight the interference-limited nature of the multi-hop architecture. We then present a hierarchical cooperation architecture that allows to overcome this interference barrier.

2.1 Model

In this section, we introduce a simple network and channel model that will be used throughout the section. This model allows us to concentrate on networks where interference is the only factor that can potentially limit performance. The model will be successively refined in the next two sections to incorporate the impacts of power and space.

2.1.1 A Random Network Model

There are n wireless nodes with transmitting and receiving capabilities, which are uniformly and independently distributed in a square of area A. Each node has an average power of P Watts and the network is allocated a total bandwidth of W Hertz around a carrier frequency of f_c, $f_c \gg W$. Every node is both a source and a destination for some

traffic. The sources and destinations are randomly paired up one-to-one without any consideration of node locations.[1] Each source has the same traffic rate $R(n)$ bits/s/Hz to send to its destination node. The aggregate throughput of the system is $T(n) = nR(n)$. (In the sequel, whenever we say that an aggregate throughput $T(n)$ is achievable, we implicitly mean that a rate of $T(n)/n$ is simultaneously achievable for all source-destination pairs in the network.) We restrict our attention to the scaling of the aggregate throughput $T(n)$ with an increasing number of nodes n. The parameters A, P and W remain constant as the number of nodes n in the network increases.

2.1.2 Channel Model

We assume that communication takes place over a flat channel and the signal received by node i at time slot m is given by

$$Y_i[m] = \sum_{k \neq i} H_{ik}[m] X_k[m] + Z_i[m], \tag{2.1}$$

where $X_k[m]$ is the signal sent by node k at time m and $Z_i[m]$ is additive white circularly symmetric Gaussian noise (AWGN) of power spectral density $N_0/2$ Watts/Hz. The complex baseband-equivalent channel gain between node i and node k at time m is given by

$$H_{ik}[m] = \sqrt{G} r_{ik}^{-\alpha/2} e^{j\theta_{ik}[m]}, \tag{2.2}$$

where r_{ik} is the distance between the nodes, $\theta_{ik}[m]$ is the random phase at time m, uniformly distributed in $[0, 2\pi)$ and $\{\theta_{ik}[m], 1 \leq i \leq n, 1 \leq k \leq n\}$ is a collection of i.i.d. random processes. The $\theta_{ik}[m]$'s and the r_{ik}'s are also assumed to be independent. The parameters G and $\alpha \geq 2$ are assumed to be constants; α is called the power path loss exponent of the environment. For example, under free-space line-of-sight propagation, Friis' formula applies and

$$|H_{ik}[m]|^2 = \frac{G_{Tx} \cdot G_{Rx}}{(4\pi r_{ik}/\lambda)^2}, \tag{2.3}$$

[1] Note that an equivalent model is to consider a network of $2n$ nodes, with n sources and n destinations placed and paired up randomly. The capacity results derived under this assumption might differ by a factor 2 from the results derived in the present monograph. However, as our main focus is the study of capacity scaling laws, this is not an issue.

so that

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

where G_{Tx} and G_{Rx} are the transmitter and receiver antenna gains, respectively, and λ is the carrier wavelength.

Note that the channel is random, depending on the location of the users and the phases. The locations are assumed to be fixed over the duration of the communication. The phases are assumed to vary in a stationary ergodic manner over time m (fast fading). The results can also be extended to the slow fading case where phases are randomly drawn from the i.i.d. uniform distribution over $[0, 2\pi)$ and then kept fixed during the time of communication. We assume that the channel gains are known at all the nodes.

Here are several comments about the model:

- The path loss model is based on a *far-field* assumption: The distance r_{ik} is assumed to be much larger than the carrier wavelength. When the distance is of the order or shorter than the carrier wavelength, the simple path loss model obviously does not hold anymore as path loss can potentially become path "gain." The reason is that near-field electromagnetics now come into play. In a scaling regime, where the area of the network is kept fixed as the number of users increases, path loss will eventually become path gain for most pairs and will be unbounded. This deficiency of the model is circumvented in the analysis as the performance depends on the signal to interference plus noise power ratio (SINR), which captures the relative strength of the paths with respect to each other and remains always finite. Signals from close-by nodes are much stronger than signals from far-away nodes. This is the only property of the model in (2.2) that we use in the analysis. This point is further elaborated at the end of Sections 2.2 and 2.3.
- The random phase model is also based on a far-field assumption: We are assuming that the nodes' separation is at a much larger spatial scale compared with the carrier wavelength, so

that the phases can be modeled as completely random and independent of the actual positions. In Section 4, we clarify how large this spatial scale should be in order for the random phase assumption to hold.

- We essentially assume a line-of-sight type environment and ignore multi-path effects. With multi-paths, there is a further randomness in the channel gains due to random constructive and destructive interference of these paths. The results of the section extend to the multi-path case.
- It is realistic to assume the variation of the phases as they vary significantly when users move a distance of the order of the carrier wavelength (fractions of a meter). The positions determine the path losses and they, on the other hand, vary over a much larger spatial scale. Hence, the positions are assumed to be fixed.

2.2 Performance of Multi-hop

In the multi-hop scheme, the packets of a source-destination pair s–d are communicated by successive point-to-point transmissions between relaying nodes. Each relay node decodes the packets from the previous relay and forwards them to the next. (Figure 2.1(a)).

We analyze the performance of multi-hop based on the following simple architecture: Let us divide the network into square cells of area A_c. Each cell contains $M = A_c n/A$ nodes on the average. Assume that each cell contains at least one node so that we can assign a node in each cell to do the relaying job. Then, packets can be transferred from one cell to the next via successive transmissions between the assigned relay nodes in each cell. Each relay decodes the packets transmitted from its neighboring cells, temporarily stores, re-encodes and forwards them to the next cell in their respective direction of transportation. Hence, the communication in the network is based on point-to-point transmissions between pairs of nodes located in neighboring cells. The SNR of these transmissions is lower bounded by

$$\mathrm{SNR}_r \geq \frac{GP}{N_0 W (\sqrt{2A_c})^{\alpha}}, \tag{2.4}$$

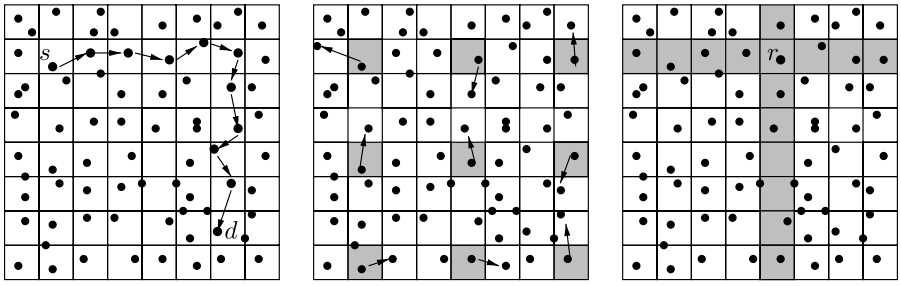

Fig. 2.1 The multi-hopping scheme: (a) The packets of a source node s are delivered to its destination node d by multi-hopping from one cell to the next. (b) A 9-TDMA scheme is employed between cells to control inter-cluster interference. The shaded cells are simultaneously active. (c) The relaying traffic at node r is originated from or destined to one of the nodes located in the shaded rectangles.

where $\sqrt{2A_c}$ is an upper bound on the distance between two nodes located in neighboring cells.

In order to increase the capacity of the architecture, it is desired to have many relaying nodes operate simultaneously inside the network. The decay of signals with distance in (2.1) allows spatially separated transmitter–receiver pairs to communicate simultaneously without creating too much interference to each other (spatial reuse). For example, in order to control interference, we can employ a simple time-division strategy (TDMA) between neighboring cells such as the 9-TDMA strategy illustrated in Figure 2.1(b). With this strategy, only the relay nodes in the shaded cells are allowed to transmit simultaneously, while nodes in the other cells are either receiving a transmission or remain inactive. This ensures that there is a guard region free of interfering transmitters around every receiving node in the network. By shifting the pattern in Figure 2.1(b) in the successive time slots, we can ensure that each relay node transmits a constant fraction 1/9 of the total time. Under such a TDMA strategy, the interference-to-noise ratio experienced by any receiver in the network can be upper bounded by

$$\mathrm{INR}_r \leq K_I \mathrm{SNR}_r \tag{2.5}$$

for a constant K_I independent of n and SNR_r. Although the proof of this fact is quite straightforward, this is a major observation that

allows spatial reuse in wireless networks. Therefore, we formally state and prove this result in Lemma 2.1 at the end of this section.

Shannon's familiar capacity formula for the point-to-point AWGN channel allows us to relate the SNR_r and INR_r to the achievable transmission rate in bits/s/Hz between two nodes in neighboring cells. Treating the interference as additional noise, the outbound rate of any relay node is given by

$$R_r = \frac{1}{9}\log\left(1 + \frac{\mathrm{SNR}_r}{1+\mathrm{INR}_r}\right) \geq \frac{1}{9}\log\left(1 + \frac{\mathrm{SNR}_r}{1+K_I\mathrm{SNR}_r}\right).$$

The factor $1/9$ accounts for the performance loss due to TDMA between neighboring clusters. When SNR_r is lower bounded by a constant independent of n, this rate is also lower bounded by a constant. As $A_c \leq A$, observe from (2.4) that for any choice of A_c, SNR_r is at least as large as $GP/(N_0W(\sqrt{2A})^\alpha)$, which is constant. Therefore, $R_r = \Theta(1)$.

Assume that the communication between each source-destination pair is relayed by following a simplistic route, first proceeding horizontally and then vertically as shown in Figure 2.1(a). Then, it is easy to observe that the relaying traffic at a particular relay node r is generated either by the source nodes located in the same horizontal slab or the destination nodes located in the same vertical slab as r. The number of nodes contained in a slab of area $\sqrt{A_cA}$ is at most $(1+\delta)\sqrt{Mn}$ w.h.p.[2] for any $\delta > 0$ by Lemma A.1. This means that the outbound rate R_r of each relay node has to be shared at most among $2(1+\delta)\sqrt{Mn}$ source-destination pairs, yielding

$$R_{MH} \geq \frac{1}{2(1+\delta)\sqrt{Mn}} R_r \tag{2.6}$$

rate per source-destination pair. Observe that reducing A_c to the nearest neighbor scale A/n, which corresponds to $M = 1$, maximizes this rate by minimizing the relaying burden. Let us define the received SNR over the typical nearest-neighbor distance $\sqrt{A/n}$ as

$$\mathrm{SNR}_s := \frac{GP}{N_0W(\sqrt{A/n})^\alpha}. \tag{2.7}$$

[2] With high probability: With probability approaching 1 when n increases. See Appendix A for a precise definition.

By choosing $A_c = A/n$, multi-hop achieves an aggregate throughput

$$T_{MH} \geq K_0 \sqrt{n} \log\left(1 + \frac{\mathrm{SNR}_s}{1 + K_I \mathrm{SNR}_s}\right), \tag{2.8}$$

where $K_0 > 0$ and $K_I > 0$ are constants independent of n. As SNR_s is lower bounded by a constant, this aggregate throughput is $\Theta(\sqrt{n})$.

There is one subtlety in the above discussion. When the cell size is reduced to A/n, our initial assumption that all cells are non-empty is violated w.h.p. However, using methods from percolation theory, it can be shown that even if some cells are empty, w.h.p. we can find-approximate horizontal and vertical paths that are composed of non-empty cells. The straight horizontal and vertical routes in Figure 2.1(a) can be replaced by these approximate straight horizontal and vertical routes and $\Omega(\sqrt{n})$ aggregate throughput scaling can be achieved (see Ref. [6] for a detailed discussion of this fact).

Is this the best scaling achievable by multi-hop? Can we improve the performance by a more careful analysis or by better optimizing the architecture? For example, the relaying burden at each node was bounded above based on the worst-case assumption that all source-destination pairs in the network are separated by the maximal $2\sqrt{A}$ Manhattan distance.[3] Or one can modify the architecture to follow straight-line routes between the source and destination pairs so as to decrease the number of hops taken by each packet. Such improvements cannot alter the scaling performance and $\Theta(\sqrt{n})$ is the best scaling achievable by the multi-hop architecture. Indeed, it can be shown that with the random network and traffic model of Section 2.1, most of the source-destination pairs in the network are separated by a distance of $\Theta(\sqrt{A})$ w.h.p. Therefore, the worst-case assumption of $2\sqrt{A}$ Manhattan distance between the source-destination pairs is order-wise precise. On the other hand, following a straight-line path between the source-destination pairs decreases the number of hops by at most a factor of $\sqrt{2}$ and does not improve the scaling performance.

[3] The Manhattan distance between two points is by definition the sum of the vertical and horizontal distances separating the two points.

To summarize, the discussion on the scaling performance of multi-hop above provides the following insights:

- Increasing the hop range $\sqrt{A_c}$ in the multi-hop architecture decreases the number of hops between source and destination nodes. However, due to interference it also decreases the number of simultaneous transmissions inside the network, the spatial reuse factor. This contention is resolved in the favor of short-range communication. Confining to nearest-neighbor transmissions maximizes the capacity of multi-hop. It can be readily observed from (2.6) that long-distance transmissions, taking $A_c = A$ or $M = n$, yield only aggregate throughput $\Theta(1)$. In this case, the architecture is reduced to simple time sharing between source-destination pairs in a round-robin fashion.
- Multi-hop is fundamentally limited by interference. Because of interference, we can accommodate $\Theta(n)$ simultaneous transmissions inside the network *only if* they are *local*, i.e., over the nearest-neighbor distance. On the other hand, the traffic demand in the network is such that we need to establish $\Theta(n)$ simultaneous *global* communications, i.e., over distances of the order of the diameter of the network. Trying to meet the global traffic demand with local transmissions leads to a fundamental relaying burden. Along these lines, Gupta and Kumar proved in [9] that conditioned on the assumption of interference being treated as noise in the network, no better aggregate throughput scaling than $O(\sqrt{n})$ can indeed be achieved in wireless networks.

We conclude the section by proving (2.5) in the following lemma. Note that the lemma holds for any choice of the cluster size A_c.

Lemma 2.1 (Spatial Reuse Lemma). Consider the 9-TDMA scheme in Figure 2.2. Let SNR_r be the received SNR at a relay node r, from its corresponding transmitter t in a neighboring cell. The node r is subject to interference from the clusters simultaneously active with t

according to the 9-TDMA scheme. For $\alpha > 2$, the interference-to-noise ratio at r satisfies

$$\mathrm{INR}_r \leq K_{I_1} \mathrm{SNR}_r,$$

where K_{I_1} is constant independent of n. When $\alpha = 2$,

$$\mathrm{INR}_r \leq K_{I_2} \mathrm{SNR}_r \log n,$$

where K_{I_2} is another constant independent of n.

Proof of Lemma 2.1. Consider a node r receiving transmission from node t. The interfering signal received by node r is given by

$$I_r = \sum_{k \in \mathcal{U}_t} H_{rk} X_k,$$

where $\mathcal{U}_t$ denotes the set of nodes transmitting simultaneously with the node t under the 9-TDMA strategy. X_k is the signal transmitted by node $k \in \mathcal{U}_t$. H_{rk} is the channel coefficient between nodes r and k given by (2.2). Note that as the interfering signals X_k transmitted from different nodes k are independent, the interference-to-noise ratio at r is given by

$$\mathrm{INR}_r = \sum_{k \in \mathcal{U}_t} \frac{GP}{N_0 W (r_{rk})^\alpha}.$$

As illustrated in Figure 2.2, the interfering cells $\mathcal{U}_t$ can be grouped based on their distance to r such that each group $\mathcal{U}_t(i)$ contains $8i$ cells or less. All the cells in group $\mathcal{U}_t(i)$ are separated by a distance larger than $(3i-2)\sqrt{A_c}$ from r for $i = 1, 2, \ldots$. Recall that A_c is the cell area. The number of such groups can be simply bounded by the total number of cells n/M in the network. Thus,

$$\begin{aligned} \mathrm{INR}_r &< \sum_{i=1}^{n/M} \sum_{k \in \mathcal{U}_t(i)} \frac{GP}{N_0 W ((3i-2)\sqrt{A_c})^\alpha} \\ &\leq \frac{GP}{N_0 W (\sqrt{A_c})^\alpha} \sum_{i=1}^{n/M} 8i \frac{1}{(3i-2)^\alpha}. \end{aligned} \tag{2.9}$$

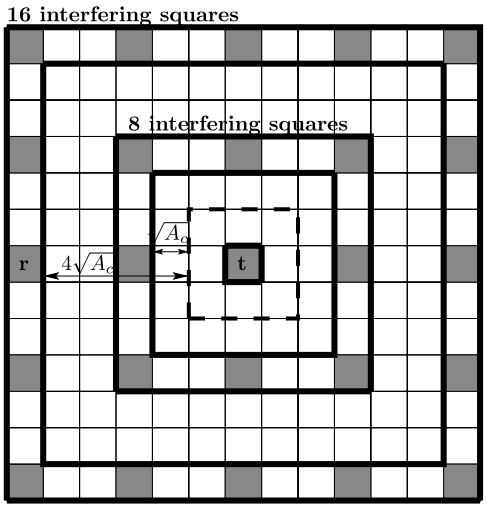

Fig. 2.2 Grouping of interfering clusters in the 9-TDMA scheme.

The last summation is convergent for $\alpha > 2$. Using this fact together with (2.4), we conclude that

$$\mathrm{INR}_r \leq K_{I_1} \mathrm{SNR}_r,$$

for a constant K_{I_1} independent of n. For $\alpha = 2$, the summation in (2.9) is of order $\log n$ and INR_r can be bounded as

$$\mathrm{INR}_r \leq K_{I_2} \log n \ \mathrm{SNR}_r,$$

where K_{I_2} is another constant independent of n. □

2.3 Hierarchical Cooperation

In this section, we present a hierarchical cooperation architecture that achieves almost linear aggregate throughput scaling under the same model. This result is stated in the following theorem.

Theorem 2.2. For any $\epsilon > 0$, there exists a constant $K_\epsilon > 0$ independent of n such that w.h.p., an aggregate throughput

$$T(n) \geq K_\epsilon n^{1-\epsilon} \tag{2.10}$$

is achievable in the network using hierarchical cooperation.

Remark 2.3. The performance in Theorem 2.2 can be achieved even if nodes use a fraction $1/n$ of their available power P. In other words, the hierarchical cooperation scheme requires only $\Theta(1/n)$ average power to achieve the scaling in Equation (2.10).

Remark 2.4. The performance in Theorem 2.2 can be achieved for any arbitrary pairing between the source and destination nodes, not necessarily a random one as defined earlier in Section 2.1.

This theorem is complemented by the one that follows, which states that there is no way to get a better capacity scaling than $O(n \log n)$ in a wireless network of size n. Therefore, in the scaling sense, the architecture in Theorem 2.2 is very close to optimal. The two theorems, Theorems 2.2 and 2.5, together establish the best possible capacity scaling in wireless networks up to logarithmic terms. We first prove Theorem 2.5 and devote the rest of the section to prove Theorem 2.2.

Theorem 2.5. The aggregate throughput achieved by any scheme is bounded above by

$$T(n) \leq K n \log n$$

w.h.p. for some constant $K > 0$ independent of n.

Proof of Theorem 2.5. One easy way to understand the result in the theorem is the following: Assume there was no interference between the transmissions inside the network. In other words, assume each source node s in the network was able to communicate to its destination node d as if they were completely on their own and the whole bandwidth of the system was granted to their exclusive use. Then the communication rate between the pair s–d is given by Shannon's capacity formula as

$$R_{sd}(n) = \log\left(1 + \frac{GP}{N_0 W r_{sd}^{\alpha}}\right), \tag{2.11}$$

where r_{sd} is the separation between the nodes s and d. By part (b) of Lemma A.1, the minimum distance between any two nodes in the network is larger than $\sqrt{A}/(n^{1+\delta})$, w.h.p, for any $\delta > 0$. Plugging this lower bound for r_{sd} in Equation (2.11), yields

$$R_{sd}(n) \leq \log\left(1 + \frac{GP}{N_0 W(\sqrt{A})^\alpha} n^{\alpha(1+\delta)}\right), \tag{2.12}$$

which is $O(\log n)$. Therefore, the aggregate throughput is $O(n \log n)$.

However, this argument is not precisely correct as the existence of other nodes can help the pair $s-d$ to enhance their communication rate beyond (2.11). A precise information-theoretical argument assumes that the only communication to establish inside the network is the one between the pair $s-d$ and all the other nodes in the network help $s-d$ to communicate. In other words, the resources of the network are granted exclusively to the pair $s-d$. Obviously, the communication rate between the pair $s-d$ under this optimistic scenario can only be larger than the rate achieved between $s-d$ in the original set-up. This optimistic communication rate between $s-d$ is, in turn, upper bounded by the total information rate that can be transfered from s to the rest of the network. This corresponds to the capacity of the single-input multiple-output (SIMO) channel between the source node s and the rest of the network. The capacity of this SIMO channel is well-known to be [30]

$$R_{\text{SIMO}} = \log\left(1 + \frac{P}{N_0 W} \sum_{\substack{i=1 \\ i \neq s}}^{n} |H_{is}|^2\right) = \log\left(1 + \frac{GP}{N_0 W} \sum_{\substack{i=1 \\ i \neq s}}^{n} \frac{1}{r_{is}^\alpha}\right).$$

Plugging the lower bound $r_{is} \geq \sqrt{A}/(n^{1+\delta})$, we observe that the resultant upper bound on $R_{sd}(n)$ differs from the one in (2.12) by a factor of n inside the logarithm. In other words, the help of the other nodes in the network can yield an increase in the SNR of the pair $s-d$ by a polynomial factor in n. This power gain translates to a constant factor gain for the communication rate. We conclude that the aggregate throughput in the original network is upper bounded by

$$T(n) \leq n \log\left(1 + \frac{GP}{N_0 W(\sqrt{A})^\alpha} n^{\alpha(1+\delta)+1}\right),$$

w.h.p for any possible communication strategy, which is again $O(n \log n)$. □

The remainder of the section is devoted to the proof of Theorem 2.2, which is based on the following lemma.

Lemma 2.6. Consider $\alpha > 2$. Assume there exists a scheme that achieves for each n, with probability at least $1 - e^{-n^{c_1}}$, an aggregate throughput

$$T(n) \geq K_1 n^b$$

for any arbitrary pairing between the source and the destination nodes (i.e., not necessarily a random pairing, as assumed earlier in Section 2.1). K_1 and c_1 are positive constants independent of n and the source-destination pairing, and $0 \leq b < 1$. Moreover, assume that the scheme is able to achieve this performance by using only P/n average power per node. Then, one can construct another scheme for this network that achieves a *higher* aggregate throughput

$$T(n) \geq K_2 n^{1/(2-b)}$$

again for any arbitrary pairing between the source and the destination nodes. $K_2 > 0$ is another constant independent of n and the pairing. The failure rate for the new scheme is upper bounded by $e^{-n^{c_2}}$ for another positive constant c_2. The per-node average power needed to achieve this higher throughput is again only P/n.

The lemma above is stated under slightly different conditions than those in Section 2.1. First, it is restricted to $\alpha > 2$, when Theorem 2.2 holds for $\alpha \geq 2$. To prove the theorem for $\alpha = 2$, one needs a slightly modified version of the lemma. The difference comes from the fact that with TDMA schemes such as the 9-TDMA scheme used in Lemma 2.1, the aggregate interference scales like $\log n$ when $\alpha = 2$, while it is constant when $\alpha > 2$. This creates a minor modification in the analysis. We concentrate only on the case $\alpha > 2$ in the remainder of this section.

The remainder of the conditions in Lemma 2.6 are somewhat more general than those in Theorem 2.2. The throughputs in the lemma

are achieved for any arbitrary pairing between the source and destination nodes, therefore in particular for a random pairing. Note that even though the pairings in Lemma 2.6 are not random but arbitrary, there is still randomness in the distribution of nodes over the network area and in the channel coefficients. The probabilities for achieving the claimed throughputs are over this remaining randomness in the problem. Note that these throughputs are achievable in any realization of the random network with exponentially small probability of failure which is stronger than our definition of high probability used in Theorem 2.2. (See Appendix A for the definition of high probability.) Finally, the required per node power in the lemma decreases to zero as P/n as n increases. As we prove next, Theorem 2.2 follows from a recursive application of Lemma 2.6, and it can be readily verified from the proof of Theorem 2.2 below that it indeed holds under these more general conditions. These observations are emphasized in Remarks 2.3 and 2.4, as this more general form of the theorem will be used in Section 3.

Proof of Theorem 2.2. Lemma 2.6 is the key step to build a hierarchical architecture and prove Theorem 2.2. As $1/(2-b) > b$ for $0 \leq b < 1$, as illustrated in Figure 2.3, the new scheme in the lemma is always better than the old one. Therefore, as soon as there is a scheme to start with, the lemma can be applied recursively, yielding a scheme that achieves better throughput scaling at each step of the recursion.

We start with a simple strategy where source-destination pairs take turns in a round-robin fashion to directly communicate with each other (TDMA). At the end of Section 2.2, we noted that such a time-division strategy corresponds to $A_c = A$ in the multi-hop formulation and achieves a throughput $\Theta(1)$, i.e., $b = 0$. This performance is independent of the source-destination pairing and the placement of nodes inside the network area. Therefore, the probability of failing to achieve the $\Theta(1)$ throughput with this strategy in a random network is 0. Moreover, as each source is only transmitting $1/n$th of the time, the average power consumed per node is $\Theta(1/n)$.

Therefore, time-division satisfies the conditions of Lemma 2.6 with $b = 0$. Starting with $b = 0$ and applying the lemma recursively h times, we get a scheme that achieves $\Theta(n^{h/(h+1)})$ aggregate throughput. Given

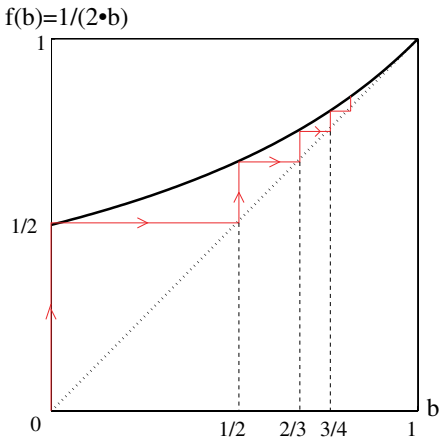

Fig. 2.3 Function $f(b) = 1/(2-b)$ representing the increase in the throughput exponent b from one level of the recursion to the next.

any $\varepsilon > 0$, we can now choose h such that $h/(h+1) \geq 1-\varepsilon$ and we get a scheme that achieves $\Theta(n^{1-\epsilon})$ aggregate throughput scaling with high probability. Alternatively, we can also start with the multi-hop strategy, which was shown to achieve a throughput $\Theta(\sqrt{n})$, i.e., $b = 1/2$, in Section 2.2 and save one step in the recursion. This concludes the proof of Theorem 2.2. □

We now sketch how the new scheme is constructed given the old scheme and provide a back-of-the-envelope analysis of the scaling law it achieves. A rigorous proof of Lemma 2.6 can be found in Ref. [26].

Proof of Lemma 2.6. The scheme that proves Lemma 2.6 is based on clustering and long-range distributed MIMO transmissions between the clusters. We divide the network into square cells of area A_c. Each cell contains a cluster of $M = A_c n/A$ nodes on the average. For the sake of simplicity, we will assume in the sequel that each cell contains exactly M nodes.[4] Let us focus now on a particular source node s and its destination node d. s sends M bits to d in the following three steps:

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

[4] In Appendix A, we show that each cell contains $\Theta(M)$ nodes with high probability. This is sufficient for proving the scaling law results in this section.

(2) These nodes together can then form a distributed transmit antenna array, sending the M bits *simultaneously* to the destination cluster where d lies,
(3) Each node in the destination cluster gets one observation from the distributed MIMO transmission. It quantizes this observation and ship it to d, which can then do joint MIMO processing of all the observations and decode the M transmitted bits from s.

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

Phase 1 Transmit Cooperation. Clusters work in parallel according to a time-division schedule similar to the 9-TDMA scheme discussed in Lemma 2.1. The TDMA scheme allows a constant fraction of the clusters to operate simultaneously. Within a cluster, each source node has to distribute its M bits among its neighbors, 1 bit for each node, such that at the end of the phase, each node has 1 bit from each of the source nodes in the same cluster. As there are M source nodes in each cluster, this gives a total traffic of exchanging $M(M-1) \sim M^2$ bits. (Recall our assumption that each node is a source for some communication request and a destination for another.) The key observation is that this is similar to the original problem of communicating between n source-destination pairs in a network of area A, but now on a smaller network of size M and area A_c. More precisely, this traffic demand of exchanging M^2 bits can be divided into M sessions. In each session, we assign M source-destination pairs to communicate their 1 bit.[5] The sessions are handled one after another. Below we will argue that we can use the scheme given in the hypothesis of the lemma to handle the traffic in each session and that it achieves an aggregate throughput

[5] As the scheme in the hypothesis of the lemma works for any pairing, the splitting of the total traffic into M sessions can be done quite arbitrarily.

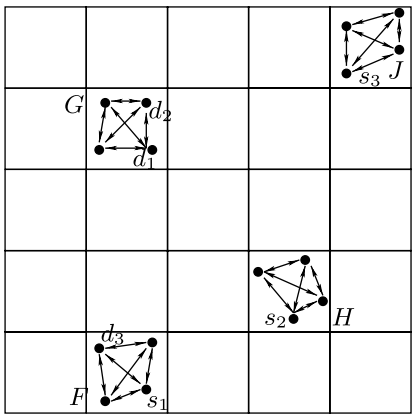

Fig. 2.4 Nodes inside clusters F, G, H and J are illustrated while exchanging packets in Phases 1 and 3. Note that in Phase 1 the exchanged packets contain the source bits whereas in Phase 3 they contain the quantized MIMO observations. Clusters work in parallel. In this Figure 2.5, we highlight three source-destination pairs s_1-d_1, s_2-d_2 and s_3-d_3, such that nodes s_1 and d_3 are located in cluster F, nodes s_2 and s_3 are located in clusters H and J, respectively, and nodes d_1 and d_2 are located in cluster G.

$\Theta(M^b)$ bits/time slot in each session. Therefore, each session can be completed in $\Theta(M^{1-b})$ time slots, and the M sessions in $\Theta(M^{2-b})$ time slots. As the time-division scheme between the clusters introduces a constant factor, this phase is completed also in $\Theta(M^{2-b})$ time slots for the whole network. (Figure 2.4).

Phase 2: Cooperative MIMO. We perform successive long-distance distributed MIMO transmissions between source-destination pairs, one transmission at a time. In each one of the MIMO transmissions, say the one between s and d, the M bits of s are simultaneously transmitted by the M nodes in its cluster to the M nodes in the cluster of d. Note that each node in the destination cluster observes a mixture of the transmitted signals from the nodes in the source cluster. In the following section, we will show that if these observations are relayed to the actual destination node d and jointly processed, the M simultaneously transmitted bits can be recovered. As the M bits from the source cluster are transmitted simultaneously, each MIMO transmission takes $\Theta(1)$ time slots. The long-distance MIMO transmissions are repeated successively for each source-destination pair; hence, we need $\Theta(n)$ time slots to complete the phase for the whole network. (Figure 2.5).

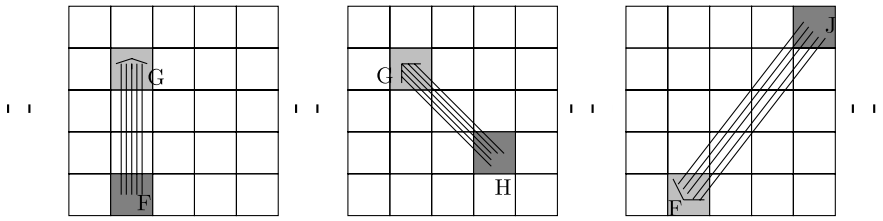

Fig. 2.5 Successive MIMO transmissions are performed between clusters. (a) depicts the MIMO transmission from cluster F to G, where packets originally belonging to s_1 are simultaneously transmitted by all nodes in F to all nodes in G, the cluster containing d_1. (b) The second MIMO transmission is from H to G, serving the s_2–d_2 pair, and (c) illustrates the MIMO transmission from cluster J to F, serving the s_3–d_3 pair.

Phase 3: Receive Cooperation. Clusters work in parallel similar to the first phase. The goal of this phase is to collect the MIMO observations from the previous phase at the actual destination nodes inside the clusters. As there are M destination nodes inside each cluster, each cluster has received M MIMO transmissions in Phase 2. Each MIMO transmission is intended for a different destination node. Thus, each node in the cluster has M received observations, one from each of the MIMO transmissions, and each observation is to be conveyed to a different destination node in its cluster. Nodes quantize each observation into fixed Q bits (independent of M and n); hence, there are now a total of QM^2 bits to exchange inside each cluster. The traffic is equivalent to the one in the first phase. Using exactly the same strategy, we conclude the phase in $\Theta(M^{2-b})$ time slots.

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

$$\frac{nM}{M^{2-b} + n + M^{2-b}} \quad \text{bits per time slot.}$$

Maximizing this throughput by choosing $M = n^{1/(2-b)}$ yields $T(n) = \Theta(n^{1/(2-b)})$ for the aggregate throughput, which is the result in

Lemma 2.6. For the sake of simplicity, we assumed that each packet contains a single bit in the above discussion. A precise description of the scheme would assume that each source node communicates M packets to its destination, each packet containing a constant number of bits. This would, in turn, scale the duration of each phase by a constant factor, but would not change the scaling law conclusion.

We made two non-trivial assumptions in the above discussion that need to be verified:

(a) The scheme in the hypothesis of Lemma 2.6 achieves M^b aggregate throughput inside the clusters in Phases 1 and 3.
(b) The distributed MIMO transmissions between two clusters achieve an aggregate rate scaling linearly in M. Equivalently, we can transmit M independent streams of constant rate from the source cluster, and the destination node is able to recover these streams from the M quantized observations it gathers at the end of Phase 3.

Moreover, Lemma 2.6 makes the following additional claims about the new scheme that we also need to verify

(c) the new scheme uses power P/n per node;
(d) the scheme works for any arbitrary pairing of the source-destination nodes;
(e) the failure probability of the new scheme is exponentially small.

We next verify the above statements:

(a) We need to verify that the scheme in the hypothesis of Lemma 2.6 can achieve M^b aggregate throughput inside the sessions in Phases 1 and 3, when a constant fraction of the clusters are operating simultaneously according to a time-division schedule, such as the 9-TDMA scheme in Lemma 2.1. Let us first argue that the scheme in the hypothesis of Lemma 2.6 requires power $P/M(A_c/A)^{\alpha/2}$ per node to achieve M^b throughput in a network of M nodes distributed on area A_c. This follows from the scale-invariance of

the model and the fact that the impact of smaller distances directly translates to larger received power: If a scheme requires power P/M when there are M nodes distributed on an area A, it will require $P/M(A_c/A)^{\alpha/2}$ power when the M nodes are distributed on an area A_c. In the later case, the distances are decreased by a factor of $\sqrt{A_c/A}$ and hence for the same transmitted power, the received power is increased by a factor of $(A/A_c)^{\alpha/2}$. More generally, according to our model in Section 2.1, a network with area A and power P/M per node is equivalent to a network of area A_c and the power per node reduced to $P/M(A_c/A)^{\alpha/2}$ instead of P/M.
Hence, we choose to operate with power $P_1 = P/M(A_c/A)^{\alpha/2}$ per node in Phases 1 and 3. The total power transmitted by a cluster is $M\,P_1$. Plugging this effective transmitted power of a cluster for P in (2.9) yields

$$\text{INR} \leq K_I \frac{GP}{N_0 W(\sqrt{A})^{\alpha}}$$

for the inter-cluster INR, which is constant. Therefore, the inter cluster interference power is of the order of the thermal noise power in the system. Simply treating it as additional, we can achieve M^b aggregate throughput simultaneously inside all clusters in Phases 1 and 3.

(b) As distributed MIMO transmission lies at the heart of the proposed architecture, we devote the following section to discuss the capacity of distributed MIMO and verify this claim.

(c) In (a), we chose to operate the scheme in the hypothesis of Lemma 2.6 with power $P_1 = P/M(A_c/A)^{\alpha/2}$ per node in Phases 1 and 3. As $M = A_c n/A$ and $\alpha \geq 2$,

$$\frac{P}{M}\left(\frac{A_c}{A}\right)^{\alpha/2} = \frac{P}{n}\left(\frac{A_c}{A}\right)^{\alpha/2-1} \leq \frac{P}{n}.$$

In the following section while verifying item (b) above, we will prove that the scheme uses power not larger P/n also in the second phase.

(d) This can be simply observed from the fact that the construction of the new scheme does not impose any constraint on the pairing of the source-destination nodes.

(e) The new scheme fails only if the cells of area A_c fail to contain order M nodes or when the old scheme fails to achieve its promised throughput inside one of the clusters in Phases 1 and 3. By Lemma A.1(b), each cell contains an order of $M = A_c n/A$ nodes with exponentially small probability of error. On the other hand, the number of times the old scheme is employed in the construction of the new scheme is polynomial in n and M (it is used in M sessions inside each of the n/M clusters) when its failure probability decreases exponentially in M. We can use the union bound to conclude that the failure probability for the new scheme is also exponentially small. □

Proving Theorem 2.2 by recursively applying Lemma 2.6, we have built a hierarchical architecture to achieve the desired throughput. At the lowest level of the hierarchy, we use the simple time-division scheme to exchange packets for cooperation among small clusters. Combining this with longer range MIMO transmissions, we get a higher throughput scheme for cooperation among nodes in larger clusters at the next level of the hierarchy. Finally, at the top level of the hierarchy, the cooperation clusters are of the order $n^{1-\epsilon}$ nodes, almost the network size, and MIMO transmissions take place on a global scale over distances of the order $\sqrt{A}$, to meet the desired traffic demands. Figure 1.3 in the introduction shows the resulting hierarchical scheme with a focus on the top two levels.

It is important to understand the aspects of the channel and the network model, which the scheme made use of in achieving the linear capacity scaling:

- The path attenuation decay law $1/r^\alpha$ ($\alpha \geq 2$) ensures that the *aggregate* signals from far-away nodes are much weaker than signals from close-by nodes. This enables (a constant fraction of) the clusters to operate simultaneously in the

first and third phases. (This is similar to the spatial reuse in multi-hop.)

- The area of the network A and the per-user power P remain constant as the number of users in the network increases. This ensures that the received SNR between every pair of nodes in the network is lower bounded by a constant $GP/(N_0W(\sqrt{2A})^\alpha)$, bounded away from zero even as the network grows. As a result, the MIMO transmissions in the second phase and therefore the overall scheme do not suffer any power limitation and can achieve linear scaling. Indeed, Remark 2.3 states that there is sufficient power to achieve linear scaling as long as $GP/(N_0W(\sqrt{2A})^\alpha)$ is $\Omega(1/n)$.
- The random i.i.d. channel phases enable full spatial multiplexing gain for the long-range MIMO transmissions, as we prove in the following section.

In the following two sections, we will investigate the optimal cooperation architectures for wireless networks when the last two conditions fail to hold.

2.4 Capacity of Distributed MIMO

In this section, we prove that the capacity of the distributed MIMO transmission between two clusters of M nodes (Figure 2.6) scales linearly in M. Roughly speaking, this means that to transmit M bits from the source cluster, we need a constant number of time slots (independent of M), which was the assumption made in Proof of Lemma 2.6. We provide two alternative proofs for this fact. We first concentrate on

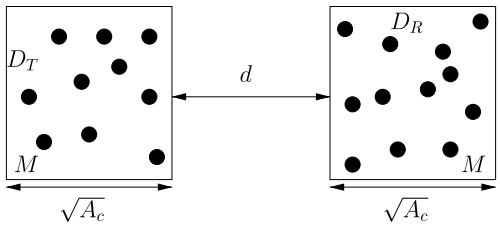

Fig. 2.6 A transmit cluster S of M nodes and a receive cluster D of M nodes separated by a distance d.

one particular decoding strategy at the final destination node. We show that using this decoding strategy the destination node is able to recover the M independent streams from the source cluster. The second proof directly starts with the well-known capacity expression for the MIMO channel from information theory and shows that it scales linearly in M. This second proof will be useful when we discuss the impact of space on the capacity of wireless networks in Section 4.

We will ignore the quantization process and assume that the observations of all the nodes inside the destination cluster are available to the destination node as they are. (The proof of the fact that the quantization process does not alter the linear scaling of the MIMO capacity can be found in Ref. [26].) Let X be an $M \times 1$ vector such that its kth entry X_k denotes the signal transmitted by node k in the source cluster S. Let Y be an $M \times 1$ vector containing the received observations of the M nodes in the destination cluster denoted by D. According to the model (2.1), X and Y are related by

$$Y[m] = H[m]X[m] + Z[m], \tag{2.13}$$

where $H[m]$ is the $M \times M$ matrix and $H_{ik}[m]$ is the channel coefficient between $k \in S$ and $i \in D$ given in (2.2). $Z[m]$ is a vector of i.i.d. circularly symmetric complex Gaussian random variables. Note that the new scheme in Lemma 2.6 is only allowed to use P/n average power per node. However, as distributed MIMO transmissions are performed successively in the second phase, each cluster transmits only a fraction M/n of the total duration of the phase and stays inactive in the rest of the time. Therefore, the nodes, when active, can transmit with power P/M per node and still satisfy their average power constraint of P/n. We assume in the sequel that $\mathbb{E}[|X_k[m]|^2] = P/M$ for all $k \in S$.

Let the distance between the two clusters S and D be r_{SD}. Observe that for any $i \in D$, $k \in S$,

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

When S and D are not neighboring clusters, we also have $r_{SD} \geq \sqrt{A_c}$. These two relations yield

$$a := \left(\frac{1}{1+2\sqrt{2}}\right)^{\alpha} \leq \left(\frac{r_{SD}}{r_{ik}}\right)^{\alpha} \leq 1, \tag{2.14}$$

when S and D are not neighboring clusters. In other words, $r_{ik}^{-\alpha} = r_{SD}^{-\alpha}\rho_{ik}$, where all ρ_{ik} lie in the interval $[a,1]$. We will make use of this fact in the sequel. When S and D are neighboring cluster, a similar result can be established by using only the nodes located on the left half of S and the nodes located in the right half of D in Figure 2.6. This ensures a separation of at least $\sqrt{A_c}$ between the transmitting and receiving nodes.

For convenience of notation, we rewrite the model in (2.13) as

$$Y[m] = bF[m]X[m] + Z[m], \tag{2.15}$$

where $b = \sqrt{G}/r_{SD}^{\alpha/2}$ and $F[m]$ is a scaled version of the channel matrix $H[m]$, namely $F_{ik}[m] = \rho_{ik}e^{j\theta_{ik}[m]}$.

2.4.1 Matched Filtering

Given the received vector $Y[m]$ and the channel matrix $F[m]$, let the destination node construct the following signal:

$$\hat{Y}[m] = F[m]^*(bF[m]X[m] + Z[m]),$$

where F^* denotes the complex conjugate transpose of the matrix F. The kth entry of this vector is given by

$$\hat{Y}_k[m] = b\sum_{i\in D}|F_{ik}[m]|^2X_k[m] + b\sum_{\substack{i\in D, l\in S,\\ l\neq k}}F_{ik}^*[m]\,F_{il}[m]X_l[m]$$
$$+ b\sum_{i\in D}F_{ik}[m]Z_i[m].$$

Let the destination node try to decode the stream $X_k[m]$ transmitted by $k \in S$ based on $\hat{Y}_k[m]$ by treating the interference from the other streams as noise. Note that the first term in the above expression corresponds to the desired signal X_k. The second term is the interference from the other streams, and the last term is the thermal noise. The noise power is upper bounded by

$$\mathbb{E}\left[\left|b\sum_{i\in D}F_{ik}[m]Z_i[m]\right|^2\right] \leq N_0W,$$

because of relation (2.14) and the fact that the random variables $Z_i[m]$, $i \in D$ at different receivers are independent. The interference power is upper bounded by

$$\mathbb{E}\left[\left|b \sum_{\substack{i\in D, l\in S,\\ l\neq k}} F_{ik}^*[m]\, F_{il}[m] X_l[m]\right|^2\right] \leq \frac{GP}{r_{SD}^{\alpha}} M,$$

because of relation (2.14) and the fact that the phases of the complex channel gains are independent for different pairs in the network. On the other hand, the received power of the desired signal is lower bounded by

$$\mathbb{E}\left[\left|b \sum_{i\in D} |F_{ik}[m]|^2 X_k[m]\right|^2\right] \geq a^2 \frac{GP}{r_{SD}^{\alpha}} M.$$

Therefore, this decoding strategy allows each transmitter $k \in S$ to transmit at a rate

$$R_k = \log\left(1 + \frac{Ma^2GP}{N_0 W r_{SD}^{\alpha} + MGP}\right).$$

As this rate is lower bounded by a constant independent of M, this implies that we can get linear scaling for the distributed MIMO transmissions in Phase 2.

2.4.2 Mutual Information

Consider the channel in (2.15). Assuming that $X_k[m]$ are i.i.d. circularly symmetric Gaussian random variables of power P/M, the mutual information of this MIMO channel, which corresponds to the number of bits/s/Hz that can be transmitted simultaneously over the channel, is given by the following expression (see Ref. [29] for detailed explanations):

$$I(X;Y,F) = \mathbb{E}\left(\log\det\left(I + \frac{\mathrm{SNR}}{M} FF^*\right)\right), \tag{2.16}$$

where $\mathrm{SNR} = GP/N_0 W r_{SD}^{\alpha}$ and the expectation is taken over the random phases. Let $\lambda_1, \ldots, \lambda_M$ denote the M eigenvalues of the matrix

$1/MFF^*$. Using the fact that the determinant of a matrix is the product of its eigenvalues, we obtain

$$I(X;Y,F) = \mathbb{E}\left(\sum_{i=1}^{M} \log(1+\mathrm{SNR}\lambda_i)\right).$$

This formula shows that at fixed SNR, the mutual information is essentially proportional to the number of non-vanishing eigenvalues of the channel matrix $1/MFF^*$ (by "non-vanishing," we mean here the eigenvalues that remain of order 1 as M increases). As a direct analysis of the asymptotic behavior of the eigenvalues is difficult to perform in the present case (because the channel matrix entries F_{ik} have different variances), we derive in the following a lower bound on the mutual information, using the Paley–Zygmund inequality of Appendix B.

Let λ be an eigenvalue picked uniformly at random among $\lambda_1,\ldots,\lambda_M$ (note that the eigenvalues $\lambda_1,\ldots,\lambda_M$ are themselves random as the matrix $1/MFF^*$ is determined by the random phases). The above mutual information can be rewritten as

$$\begin{aligned} I(X;Y,F) &= M\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}(\log(1+\mathrm{SNR}\lambda_i)), && (2.17)\\ &= M\,\mathbb{E}(\log(1+\mathrm{SNR}\,\lambda)), && (2.18)\\ &\geq M\log(1+\mathrm{SNR}\,t)\mathbb{P}(\lambda > t) \end{aligned}$$

for any $t \geq 0$. Equation (2.17) follows by linearity of the expectation. In order to obtain (2.18), we recognize that the resultant expression in Equation (2.17) contains an averaging over the eigenvalues. The expectation in Equation (2.18) is over the random variable λ. By the Paley–Zygmund inequality in Lemma B.1 for a non-negative random variable λ and a constant t such that $0 \leq t < \mathbb{E}(\lambda)$, we have

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

We therefore obtain

$$I(X;Y,F) \geq M\log(1+\mathrm{SNR}t)\frac{(\mathbb{E}(\lambda)-t)^2}{\mathbb{E}(\lambda^2)}$$

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

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

$$= \frac{1}{M^2}\sum_{i,k=1}^{M}\rho_{ik}^2 \geq a^2.$$

The first equality follows from the fact that $\sum_{i=1}^{M}\lambda_i = \mathrm{Tr}(1/MFF^*)$ and that the expected value of the randomly and uniformly chosen eigenvalue λ is given by $1/M\sum_{i=1}^{M}\lambda_i$. The rest simply follows by evaluating the resultant expectation with respect to the independent distribution across different channel gains. For $\mathbb{E}(\lambda^2)$, by using the fact $\sum_{i=1}^{M}\lambda_i^2 = \mathrm{Tr}(1/MFF^*1/MFF^*)$ we obtain

$$\mathbb{E}(\lambda^2) = \frac{1}{M}\mathbb{E}\left(\mathrm{Tr}\left(\frac{1}{M^2}FF^*FF^*\right)\right) = \frac{1}{M^3}\sum_{iklm=1}^{M}\mathbb{E}(F_{ik}\overline{F_{lk}}F_{lm}\overline{F_{im}}),$$

$$\leq \frac{2}{M^3}\sum_{ikm=1}^{M}\mathbb{E}(|F_{ik}|^2)\mathbb{E}(|F_{im}|^2) = \frac{2}{M^3}\sum_{ikm=1}^{M}\rho_{ik}^2\rho_{im}^2 \leq 2;$$

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

$$I(X;Y,F) \geq M\log(1+\mathrm{SNR}t)\frac{(a^2-t)^2}{2}. \tag{2.19}$$

Choosing, for example, $t = a/2$ shows that $I(X;Y,F)$ grows at least linearly with M. The exact statement and the proof of the Paley–Zygmund inequality is given in Appendix B.

3

Power

In this section, we study the impact of power on the capacity of wireless networks. For this purpose, we first extend the scaling law formulation in Section 2.1 to a multi-parameter scaling-law problem. The way we scaled the parameters of the network in the earlier section (increase the number of users as the area of the network, the per-user power and the bandwidth of the system remain fixed) results in a network where all the pairwise channels are strong (of high SNR). The goal of the current section is to understand optimal architectures for wireless networks where this is not necessarily the case. For this purpose, we study different scalings of the system parameters in this section, which uncovers power-limited operating regimes in wireless networks. Indeed, to obtain a complete picture of the impact of power on capacity, we need to study all possible limits with respect to the key parameters. This leads to a multi-parameter scaling-law problem, which is formulated in Section 3.1.

Using this new formulation, we first evaluate the performance of the two architectures discussed in the previous section, multi-hop and hierarchical cooperation, in power-limited networks. In Section 3.3, we present a hybrid architecture that combines these two strategies and

we show that this new architecture performs strictly better than both of the earlier architectures in most interesting parameter ranges. In Section 3.5, we derive a multi-parameter information-theoretic upper bound on the scaling of the network capacity. This upper bound shows that the three architectures mentioned so far, multi-hop, hierarchical cooperation and a combination of the two, are sufficient for achieving the capacity of all wireless networks, as far as scaling is concerned. At the same time, none of these strategies can be left out. That is, there is a regime where each of these architectures performs strictly better than the others.

3.1 A Multi-Parameter Scaling-Law Problem

Shannon's classical capacity formula

$$C_{\text{AWGN}}(W, P_r/N_0) = W \log_2 \left(1 + \frac{P_r}{N_0 W}\right) \text{ bits/s} \tag{3.1}$$

of a point-to-point additive white Gaussian noise (AWGN) channel with bandwidth W Hz, received power P_r Watts, and white noise with power spectral density $N_0/2$ Watts/Hz plays a central role in communication system design. The formula not only quantifies exactly the performance limit of communication in terms of system parameters, but perhaps more importantly also identifies two fundamentally different operating regimes . The operating regime of the channel is determined by the key parameter: the SNR defined as

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

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

formula for the capacity

$$C_{\mathrm{AWGN}}(W, P_r/N_0) \propto \begin{cases} W, & \text{if SNR} \gg 0\,\text{dB}, \\ P_r/N_0, & \text{if SNR} \ll 0\,\text{dB}. \end{cases} \tag{3.2}$$

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

Now, imagine the capacity formula (3.1) was not at our disposal and we were interested in finding (3.2) that approximates the dependence of the capacity on the two resources in the channel and identifies the two qualitatively different operating regimes depending on SNR. The approximation (3.2) can be obtained by studying the interplay between the two resources, the bandwidth and the power, as a scaling law problem. Suppose P_r/N_0 and W are coupled to each other via a parametric formula, $P_r/N_0 = W_0 m^{\gamma_1}$ and $W = W_1 m^{\gamma_2}$ with γ_1, γ_2 fixed real numbers and m a dummy parameter. W_0 and W_1 are positive constants of appropriate units. Assume further that for any γ_1, γ_2, we are able to characterize the scaling exponent in m of the spectral efficiency $\rho_{\mathrm{AWGN}} = C_{\mathrm{AWGN}}/W$ in bits/s/Hz

$$e_{\mathrm{AWGN}}(\gamma_1, \gamma_2) := \lim_{m \to \infty} \frac{\log \rho_{\mathrm{AWGN}}(\gamma_1, \gamma_2)}{\log m}. \tag{3.3}$$

For the AWGN channel, we would obtain

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

This can be expressed in a simpler form as

$$e_{\mathrm{AWGN}}(\gamma) = \begin{cases} 0, & \text{if } \gamma \geq 0, \\ \gamma, & \text{if } \gamma < 0. \end{cases} \tag{3.4}$$

if we define $\gamma = \gamma_1 - \gamma_2$ and $\text{SNR} = P_r/N_0 W = m^{\gamma}$. (From now on, we ignore the constants W_0, W_1 that are required for matching the units in the parametric formula above, but do not change the scaling law.) The characterization of the scaling exponent in (3.4) can be used to deduce the approximation (3.2) for the capacity: Note that for large m, $\gamma > 0$ corresponds to $\text{SNR} \gg 0\,\text{dB}$. In this case, the spectral efficiency (C_{AWGN}/W) in (3.4) is constant, i.e., does not exhibit

a significant dependence on either W or P_r. Equivalently, the capacity in bits/s is linear in W as given in (3.2). Therefore, the scaling law problem (3.3) when characterized for any γ_1, γ_2 can be used as a tool to discover the operating regimes of the AWGN channel and obtain an approximate characterization of its capacity. We next define a similar scaling law problem for wireless networks, the solution of which leads to an analogous characterization of the capacity of large wireless networks.

In Section 2, we were interested in characterizing the scaling exponent of the aggregate throughput $T(n) = nR(n)$ with the number of nodes n, when the parameters A, P and W remained constant as n increases. Equivalently, A, P and W were coupled with n as $A = V_0\, n^0$, $P = V_1\, n^0$, $W = V_2\, n^0$, where V_0, V_1 and V_2 are constants with appropriate units. This particular limit allowed us to focus exclusively on the impact of interference in wireless networks, as it results in a high SNR regime for the whole network: even the received SNR between the most far away pairs in the network, given by $P/N_0W(\sqrt{A})^\alpha$, remains lower bounded by a constant when n increases. This models the scenario where all users have high SNR connections with each other. This is a practically relevant case as illustrated in Example 1.1. As can be observed directly from the SNR expression $\frac{P}{N_0W(\sqrt{A})^\alpha}$, a network can be in this high SNR regime due to a number of reasons: a large number of users can be distributed on a relatively small area, the power available at the wireless users can be large or the available bandwidth can be relatively small.

We now want to address networks that can potentially suffer from power limitation. For this purpose, we can consider another limit that now results in vanishing received SNR between some pairs of nodes in the network as n increases. For example, $A = V_0\, n^1$, $P = V_1 n^0$, $W = V_2 n^0$ is one such limit, which yields a low SNR regime for most pairwise channels inside the network. It can be readily verified that the SNR between all pairs inside the network either decreases to zero with increasing n or remains upper bounded by a constant.[1] However, considering one arbitrary limit can miss out much of the interesting

[1] We interpret a channel in both high and low SNR, if the SNR does not depend on n.

parameter space, especially because, as we will see later in this section, there are multiple power-limited regimes in wireless networks. A single limit can capture only one of them. Moreover, this particular limit $A = V_0 n^1,\ P = V_1 n^0,\ W = V_2 n^0$, usually referred to as the extended scaling in the literature, is not the most interesting one: It does not allow us to study the common scenario where the channels between nearby nodes are in the high SNR regime, while those between far-away nodes are in the low SNR regime.

In this section, we want to study all possible interplay between A, P and W. These are three independent parameters of the network, each of which can take on a wide range of values in practice. In complete analogy to the AWGN case, we formulate the interplay as a scaling law problem, focusing on the large n limit. In the most general sense, we let $A = n^{\beta_1}$, $P = n^{\beta_2}$, $W = n^{\beta_3}$ and identify the scaling exponent

$$e(\beta_1, \beta_2, \beta_3) := \lim_{n \to \infty} \frac{\log T(n, \beta_1, \beta_2, \beta_3)}{\log n} \tag{3.5}$$

of the aggregate throughput $T(n) = nR(n)$ in bits/s/Hz, for any $\beta_1, \beta_2, \beta_3$. (Note that the actual aggregate capacity of the network in bits/s is the bandwidth times $T(n)$.)

In parallel to the point-to-point AWGN case, the scaling law problem in (3.5) can be expressed in a simpler form. For the channel model in (2.2), we will see that the aggregate throughput (expressed in bits/s/Hz) depends on A, P and W only through a single SNR parameter. As opposed to the point-to-point case, there are many SNR parameters in a network corresponding to channels between different pairs of nodes. We can take any of these different SNR parameters as a reference. Here, without loss of generality, we choose to work with the received SNR over the typical nearest-neighbor distance in the network, denoted by SNR_s. Already defined in (2.7), it corresponds to the received SNR in a point-to-point transmission over the typical nearest-neighbor distance $\sqrt{A/n}$ in the network.[2]

[2] Note that from SNR_s and n, we can determine the SNR for communicating at any other scale in the network. For example, the SNR between the most far-away pairs in the network, typically separated by a distance $\Theta(\sqrt{A})$, is $n^{-\alpha/2}\mathrm{SNR}_s$ as the diameter is $\sqrt{n}$ times the nearest-neighbor distance.

Therefore, the scaling law problem in (3.5) can be equivalently stated as characterizing the scaling exponent

$$e(\beta) := \lim_{n\to\infty} \frac{\log T(n,\beta)}{\log n} \tag{3.6}$$

of the aggregate throughput $T(n)$ for any real β where $\mathrm{SNR}_s = n^\beta$. In the setting of Section 2, $\mathrm{SNR}_s = n^{\alpha/2}$, where α is the power path loss exponent of the environment in (2.2). We have seen in this case that

$$\lim_{n\to\infty} \frac{\log T(n)}{\log n} = 1,$$

hence, $e(\alpha/2) = 1$. The main technical result in the present section is the characterization of $e(\beta)$ for any real β.

3.2 Multi-hop and Hierarchical Cooperation in Power-Limited Networks

In this section, we evaluate the performance of multi-hop and hierarchical cooperation in power-limited wireless networks.

3.2.1 Multi-hop

The aggregate throughput of multi-hop is given in (2.8)

$$T_{MH} \geq K_0\sqrt{n}\log\left(1 + \frac{\mathrm{SNR}_s}{1 + K_I\mathrm{SNR}_s}\right),$$

where K_0 and K_I are constants. As multi-hop is based on nearest-neighbor transmissions, not surprisingly, its aggregate throughput is determined by SNR_s. When $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$, or equivalently when $\mathrm{SNR}_s = n^\beta$ and $\beta > 0$, $T_{MH} = \Theta(\sqrt{n})$. This was the case in the previous section. The multi-hop architecture becomes power-limited when the nearest-neighbor channels are in the low-SNR regime. In this case, $T_{MH} = \Theta(\sqrt{n}\mathrm{SNR}_s)$. To summarize,

$$T_{MH} = \begin{cases} \Theta(\sqrt{n}), & \text{if } \mathrm{SNR}_s \gg 0\,\mathrm{dB}, \\ \Theta(\sqrt{n}\mathrm{SNR}_s), & \text{if } \mathrm{SNR}_s \ll 0\,\mathrm{dB}, \end{cases} \tag{3.7}$$

or in terms of scaling exponent,

$$e_{MH}(\beta) = \begin{cases} 1/2, & \text{if } \beta \geq 0, \\ 1/2 + \beta, & \text{if } \beta < 0. \end{cases} \tag{3.8}$$

3.2.2 Bursty Hierarchical Cooperation

From Theorem 2.10 and Remark 2.3 in Section 2, the hierarchical cooperation architecture achieves an aggregate throughput $\Theta(n^{1-\varepsilon})$ for any $\varepsilon > 0$ when the power available at the nodes is $\Omega(1/n)$ (and also $W = \Theta(1)$ and $A = \Theta(1)$). As the impact of power is to determine the received SNR's in the network, this power requirement can be equivalently stated in terms of received SNR. Let us take the nearest-neighbor SNR defined in (2.7) as a reference. It can be easily verified from the definition of SNR_s and the requirement $P = \Omega(1/n)$ when $W = \Theta(1)$ and $A = \Theta(1)$ that hierarchical cooperation achieves $\Theta(n^{1-\varepsilon})$ scaling if $\mathrm{SNR}_s = \Omega(n^{\alpha/2-1})$. Note that stating the power requirement of the scheme in terms of SNR is more informative; it says that whether linear scaling can be achieved in a network or not is determined by the joint impact of P, A and W on SNR (and not their individual scalings). Now, what is the performance of hierarchical cooperation in networks where the parameters P, A and W are such that $\mathrm{SNR}_s = O(n^{\alpha/2-1})$?

In such networks, we can consider a simple "bursty" modification of the hierarchical cooperation architecture, which runs the hierarchical scheme as it is during a fraction

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

of the time with elevated power $P/(n^{1-\alpha/2}\,\mathrm{SNR}_s)$ per node and remains silent for the rest of the time. (Note that as $\mathrm{SNR}_s = O(n^{\alpha/2-1})$, $n^{1-\alpha/2}\mathrm{SNR}_s \leq 1$ for large n.) This bursty strategy consumes

$$\frac{P}{n^{1-\alpha/2}\mathrm{SNR}_s} \cdot n^{1-\alpha/2}\mathrm{SNR}_s = P$$

power on the average and achieves an aggregate throughput scaling

$$n^{1-\alpha/2}\mathrm{SNR}_s \cdot n^{1-\varepsilon}.$$

Note that during operation, the effective nearest-neighbor SNR is increased by a factor of $1/n^{1-\alpha/2}\mathrm{SNR}_s$ as the transmit power is increased by the same factor. Therefore, $\mathrm{SNR}_s = \Theta(n^{\alpha/2-1})$ and the scheme achieves $\Theta(n^{1-\varepsilon})$ aggregate throughput. However, as the scheme operates only a fraction of the time, the effective aggregate throughput is $n^{1-\alpha/2}\mathrm{SNR}_s$ times this.

In terms of the scaling exponent of the aggregate throughput, hierarchical cooperation together with this bursty modification achieves

$$e_{HC}(\beta) = \begin{cases} 1, & \text{if } \beta \geq \alpha/2 - 1, \\ 2 - \alpha/2 + \beta, & \text{if } \beta < \alpha/2 - 1, \end{cases}$$

when $\mathrm{SNR}_s = n^\beta$.

Note that this "bursty" transmission strategy has a high peak-to-average power ratio. However, although we consider bursty transmissions in time in the above discussion, such burstiness can just as well be implemented over frequency with only a fraction of the total bandwidth W used. For example, this can be implemented in an OFDM system, using a subset of the sub-carriers at any one time, but putting more energy in the active sub-carriers. This way, the peak power remains constant over time.

In the above discussion, the quantity $n^{1-\alpha/2}\mathrm{SNR}_s$ stands as a major parameter. Depending on whether this quantity is larger or smaller than $\Theta(1)$, the hierarchical cooperation architecture is power-limited or not. There is a physical-interpretation to this quantity as the long distance SNR in the network. Let us define

$$\mathrm{SNR}_l := n^{1-\alpha/2}\mathrm{SNR}_s, \tag{3.9}$$

or equivalently from (2.7)

$$\mathrm{SNR}_l = n\frac{GP}{N_0 W(\sqrt{A})^\alpha}. \tag{3.10}$$

This quantity corresponds to n *times* the received SNR of a transmitter–receiver pair separated by a distance equal to the diameter of the network. There are $\Theta(n)$ nodes located at distance $\Theta(\sqrt{A})$ to any given node in the network. Hence, n times the SNR between the most far-away nodes can be interpreted as the total SNR that can be transferred to a node over this largest spatial scale. Similarly, the short-distance SNR in (2.7) can be interpreted as the total SNR that can be transfered to a node over the nearest-neighbor scale. However, as nodes have only a constant number of nearest neighbors, the short-distance SNR is simply the SNR of a nearest-neighbor pair. Note that when $\alpha \geq 2$, the long-distance SNR is always smaller than or equal to the short-distance SNR in the network.

In terms of this new SNR parameter, the performance of hierarchical cooperation can be expressed as

$$T_{HC} = \begin{cases} \Theta(n^{1-\epsilon}), & \text{if } \mathrm{SNR}_l \gg 0\,\mathrm{dB}, \\ \Theta(n^{1-\epsilon}\mathrm{SNR}_l), & \text{if } \mathrm{SNR}_l \ll 0\,\mathrm{dB}. \end{cases} \tag{3.11}$$

Therefore, roughly speaking, $T_{HC} = n^{1-\varepsilon}\log(1+\mathrm{SNR}_l)$. This corresponds to the capacity of a MIMO transmission between two clusters of antennas, each of size $\Theta(n^{1-\epsilon})$ and separated by a distance $\Theta(\sqrt{A})$. The backbone of the hierarchical cooperation architecture is such MIMO transmissions performed at the highest level of the hierarchy. Therefore, the performance is roughly given by the capacity of these MIMO transmissions.

Comparing the performances of multi-hop and hierarchical cooperation in (3.7) and (3.11), respectively, together with the relation $\mathrm{SNR}_l = n^{1-\alpha/2}\mathrm{SNR}_s$, we observe that the network experiences power limitation if $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$. In such power-limited networks, hierarchical cooperation performs better than multi-hop when $2 \le \alpha \le 3$; observe that the second line of (3.11) is always larger than the second line of (3.7) in this case. Signal power decays slowly with distance when $2 \le \alpha \le 3$, and hierarchical cooperation yields maximal received power by collecting the received signals of a large number of nodes. Therefore, it performs better than multi-hop. When $\alpha > 3$, signal power decays fast with distance; hence, long-distance communications are not preferable. Multi-hop performs better than hierarchical cooperation in this case; the second line of (3.7) is larger than the second line of (3.11).

3.3 A Hybrid Architecture: MIMO — Multi-hop

Is this the best performance we can get in power-limited wireless networks, i.e., when $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$? We next show that a hybrid architecture combining hierarchical cooperation with multi-hop performs significantly better than either of these strategies alone when $\alpha > 3$ and $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$. Note that as $\mathrm{SNR}_s = n^{\alpha/2-1}\mathrm{SNR}_l$, there is a wide range of parameters where $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$, while $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$. This corresponds to the heterogeneous case where the short-range links are strong (of high SNR) and the long-range links are weak (of low SNR) in a network.

Theorem 3.1. Let $\alpha > 2$, $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$ and $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$. For any $\epsilon > 0$, there exists a constant $K_3 > 0$ independent of n such that w.h.p., an aggregate throughput

$$T_{\mathrm{MIMO-MH}} \geq K_3 \sqrt{n} \mathrm{SNR}_s^{1/(\alpha-2)-\epsilon}$$

is achievable in the network, using a hybrid architecture combining hierarchical cooperation with multi-hop.

In terms of scaling exponent, this result can be stated as

$$e_{\mathrm{MIMO-MH}}(\beta) = \frac{1}{2} + \frac{\beta}{\alpha - 2}, \quad \text{if } \alpha > 2, \quad \beta \geq 0 \quad \text{and} \quad \beta < \alpha/2 - 1.$$

Note that for $\beta \geq 0$, this scaling exponent is larger than $1/2$, the scaling exponent achieved by pure multi-hop. In particular, when $\alpha > 3$, it is also better than the scaling exponent achieved by pure hierarchical cooperation, which performs worse than multi-hop in this case, as pointed out earlier.

We describe in detail how the hybrid architecture operates in the proof of Theorem 3.1. On the global scale, this hybrid architecture is similar to multi-hop. The network is divided into cells and the packets of each source-destination pair are transferred by hopping from one cell to the next. At each hop, the packets are decoded and then re-encoded for the next hop. The architecture differs from multi-hop by the fact that each hop is performed via cooperative MIMO transmission assisted by hierarchical cooperation. The architecture is illustrated in Figure 1.4.

Proof of Theorem 3.1. Let us divide the network into square cells of area A_c. Let $M = A_c n/A$ be the average number of nodes contained in each cell. Below, we argue more precisely that for our particular choice of A_c, Lemma A.1(a) ensures that there are $\Theta(M)$ nodes in all cells w.h.p. We relay the packets of the source-destination pairs by hopping from one cell to the next and perform each hop by distributed MIMO transmissions. As in the case of pure multi-hopping, we follow a simplistic route between the source-destination pairs by first relaying the

packets horizontally and then vertically, as shown in Figure 1.4. Hence, the relaying burden imposed on a given cell is due to the source nodes that lie in its horizontal slab and destination nodes that lie in its vertical slab. The number of nodes contained in a slab of area $\sqrt{A_c}A$ is at most $(1+\delta)\sqrt{Mn}$ w.h.p. for any $\delta > 0$ by Lemma A.1(a). Hence, there can be at most $O(\sqrt{Mn})$ source-destination routes that pass through a given cell. Let us arbitrarily associate the source-destination pairs whose routes pass through a given cell with one of the M nodes in this cell, such that each node is associated with at most $O(\sqrt{n/M})$ source-destination pairs. The only rule that we need to respect while doing this association is that if a source-destination route starts or ends in a certain cell, then the node associated with this source-destination pair in this cell should naturally be its source or destination node, respectively. The nodes associated with a source-destination pair will act as relays for this source-destination pair during the multi-hop operation. They will decode, temporarily store and forward the packets of this source-destination pair. At each hop, the packets of the source-destination pair will be transferred from its relay node in one cell to its relay node in the next cell via distributed MIMO transmissions.

Note that the total relaying traffic departing from a given cell is composed of $O(\sqrt{Mn})$ point-to-point links between the nodes in this cell and the nodes in its four neighboring cells, such that each node is source for $O(\sqrt{n/M})$ links. This traffic can be organized into $O(\sqrt{n/M})$ sessions such that in each session we assign M links with source nodes in one cell and destination nodes in a neighboring cell to relay their packets. Note that these two neighboring cells together with the traffic in each session can be viewed as a small wireless network of $2M$ nodes randomly and uniformly distributed on a rectangular area $2\sqrt{A_c} \times \sqrt{A_c}$ and paired to M source-destination pairs. (Consider, for example, the two cells highlighted in Figure 1.4.) Assume the long-distance SNR in this small network, $\mathrm{SNR}_l(A_c)$, is $\gg 0\,\mathrm{dB}$, where

$$\mathrm{SNR}_l(A_c) = M\frac{GP}{N_0 W(\sqrt{A_c})^\alpha} = M^{1-\alpha/2}\mathrm{SNR}_s. \tag{3.12}$$

If we use hierarchical cooperation to establish the M links in this small network, we get an aggregate throughput $\Theta(M^{1-\epsilon})$ by (3.11). As $\mathrm{SNR}_l(A_c) \gg 0\,\mathrm{dB}$, there is no power limitation for hierarchical cooperation in this small network. (Note that by Remark 2.4, $\Theta(M^{1-\epsilon})$ scaling can be achieved for any arbitrary source-destination pairing.) This corresponds to $M^{-\varepsilon}$ rate for each of the M links involved. As we have to time-share between the $O(\sqrt{n/M})$ sessions, with this strategy the cell can provide an outbound relaying rate of $\Theta(\sqrt{M}M^{-\epsilon}/\sqrt{n})$ to each of the source-destination pairs that are routed through the cell.

To conclude that the hybrid architecture achieves a rate $\Theta(\sqrt{M}M^{-\epsilon}/\sqrt{n})$ per source-destination pair, we should ensure that we can provide this rate simultaneously at all hops. A time-division strategy between the transmitting cells, such as the 9-TDMA scheme used in Section 2.2, can be used to control the inter-cell interference, so that when treated as additional noise, the inter-cell interference does not degrade significantly the hop capacity derived above. Such a TDMA strategy also ensures that each cell is active a constant fraction of the time; hence, the overhead introduced by the TDMA strategy does not alter the scaling.

We conclude that the aggregate throughput achieved by the hybrid architecture is given by

$$T_{\mathrm{MIMO-MH}} = \Theta(\sqrt{n}M^{1/2-\epsilon}). \tag{3.13}$$

We therefore see that combining multi-hop with hierarchical cooperation provides an $\sqrt{M}$-fold gain for the aggregate throughput as compared with pure multi-hop, which corresponds to $M = 1$ in the above discussion. Choosing larger M in (3.13) yields a larger aggregate throughput, as it increases the hop capacity. Note that if we could choose $M = n$, we could get linear scaling, in which case the scheme reduces to pure hierarchical cooperation. However, as $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$, the assumption $\mathrm{SNR}_l(A_c) \gg 0\,\mathrm{dB}$ we have made earlier is not satisfied for $A_c = A$; hence, M cannot be as large as n. From (3.12), we see that the largest cluster size that satisfies the condition $\mathrm{SNR}(A_c) \gg 0\,\mathrm{dB}$ is given by

$$M = \mathrm{SNR}_s^{1/(\alpha/2-1)} \quad \text{and} \quad A_c = (A/n)\,\mathrm{SNR}_s^{1/(\alpha/2-1)}. \tag{3.14}$$

This is the largest geographical scale in the network over which the power limitation is not felt. Any larger cluster size increases the relaying burden without increasing the hop capacity. Combining (3.14) and (3.13) completes the proof of Theorem 3.1. □

Remark 3.2. Note that the above strategy is a decode-and-forward strategy, in the sense that after each hop, clusters perform the entire decoding of the received message (via local hierarchical cooperation), before retransmitting it further. One could then ask whether pure amplify-and-forward MIMO transmissions between clusters would not be easier to perform in this case. The problem is that the throughput of such a scheme degrades with increasing system size, because of two problems: noise amplification at each hope is the first one; second the end-to-end MIMO channel matrix looses degrees of freedom over multiple hops.

3.4 Operating Regimes of Large Wireless Networks

Combining the performances of the three architectures discussed so far, pure hierarchical cooperation, pure multi-hop and the hybrid combination of the two in Theorem 3.1, we obtain the following lower bound for the scaling exponent of wireless networks:

$$e(\alpha,\beta) \geq \begin{cases} 1, & \beta \geq \alpha/2 - 1, & \text{Hierarchical cooperation,} \\ 2 - \dfrac{\alpha}{2} + \beta, & \beta < \alpha/2 - 1, \text{ and } 2 \leq \alpha < 3, & \text{Bursty HC,} \\ \dfrac{1}{2} + \beta, & \beta \leq 0 \text{ and } \alpha \geq 3, & \text{Multi-hop,} \\ \dfrac{1}{2} + \dfrac{\beta}{(\alpha-2)}, & 0 < \beta < \alpha/2 - 1 \text{ and } \alpha \geq 3, & \text{MIMO multi-hop.} \end{cases} \tag{3.15}$$

Note that we shifted notation from $e(\beta)$ to $e(\alpha,\beta)$, in order to emphasize the explicit dependence on α. In Section 3.5, we prove that this lower bound on $e(\alpha,\beta)$ is tight by deriving a matching

information-theoretic upper bound on $e(\alpha, \beta)$. This implies that the architectures in (3.15) yield the best possible scaling exponent in the corresponding regimes; therefore, these are scaling optimal architectures for wireless networks.

In parallel to the point-to-point case discussed in Section 3.1, the characterization of the scaling exponent in (3.15) allows one to obtain an approximation of the capacity of wireless networks, which identifies the dependence of the capacity to major system parameters. In order to obtain a formula similar to (3.2), let us define P_r as the received power from a node at the typical nearest-neighbor distance. The earlier defined quantities SNR_s and SNR_l can be expressed in terms of this quantity as

$$\mathrm{SNR}_s = \frac{P_r}{N_0 W} \quad \text{and} \quad \mathrm{SNR}_l = \frac{n^{1-\alpha/2} P_r}{N_0 W}.$$

Then, the total capacity C of large wireless networks, in bits/s, is approximately given by

$$C \propto \begin{cases} nW, & \text{if } \mathrm{SNR}_l \gg 0\,\mathrm{dB}, \\ n^{2-\alpha/2} P_r / N_0, & \text{if } \mathrm{SNR}_l \ll 0\,\mathrm{dB} \\ & \quad \text{and } 2 \le \alpha \le 3, \\ \sqrt{n} P_r / N_0, & \text{if } \mathrm{SNR}_s \ll 0\,\mathrm{dB} \quad \text{and} \quad \alpha > 3, \\ \sqrt{n} W^{\frac{\alpha-3}{\alpha-2}} (P_r/N_0)^{1/(\alpha-2)}, & \text{if } \mathrm{SNR}_l \ll 0\,\mathrm{dB}, \mathrm{SNR}_s \gg 0\,\mathrm{dB} \\ & \quad \text{and } \alpha > 3. \end{cases} \tag{3.16}$$

This approximation identifies four qualitatively different regimes in wireless networks, depending on the two SNR parameters we defined earlier: the short-distance SNR and the long-distance SNR.

The first regime in (3.16) is degrees of freedom-limited regime. The bandwidth and the number of nodes in the network together constitute the available degrees of freedom in the system. In this regime, the network does not face any power limitation, as even the long distance SNR in the network is large. Thus, long distance communication is feasible and good communication schemes should exploit this feasibility. On the other hand, the network is degrees of freedom limited; hence, good communication schemes for this regime should also achieve the full degrees

of freedom in the system. These are the properties of the hierarchical cooperation architecture: The communication in the network is done via cooperative MIMO transmissions between large clusters of nodes (of size almost of order n) and at distance of the order of the diameter of the network. The performance of the MIMO transmissions is linear in the number of nodes, implying that interference limitation is removed by cooperation, and full degrees of freedom are achieved, at least as far as scaling is concerned.

In all the other regimes, the long-distance received SNR is less than 0 dB. Hence, the network is power-limited and the transfer of power becomes important in determining performance. In the second regime, i.e., when $\alpha \leq 3$, signal power decays slowly with distance and the total power transfer is maximized by global cooperation. Cooperative MIMO communication not only achieves the full degrees of freedom in the system, but it also provides a power gain by combining signals received at different nodes. This power gain becomes important in this regime. Note that this is a power-limited regime; hence, the performance depends critically on the available power, but not so much on the bandwidth.

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

The most interesting regime and the one that is most counterintuitive, given our understanding of the point-to-point AWGN

channel, is the fourth regime, when $\alpha > 3$ and $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$; however, $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$. Note that as $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$, this is still a power-limited regime and optimal schemes should transfer power efficiently across the network. However, $\mathrm{SNR}_s \gg 0\,\mathrm{dB}$; hence, the nearest-neighbor transmissions are now bandwidth-limited and not power-efficient in translating the power gain into capacity gain. There is potential for increasing the throughput by spatially multiplexing transmission via cooperation within clusters of nodes and performing distributed MIMO. Yet, the clusters cannot be as large as the size of the network, as $\mathrm{SNR}_l \ll 0\,\mathrm{dB}$ and power attenuates rapidly for $\alpha > 3$. The exact cooperation scale is dictated by the exact amount of power available and the power path loss exponent.

It turns out that the optimal scheme in this regime is to cooperate hierarchically within clusters of an intermediate size, perform MIMO transmission between adjacent clusters and then multi-hop across several clusters in order to reach the final destination. The optimal cluster size is chosen such that the received SNR in the MIMO transmissions is at 0 dB. Any smaller cluster size results in power inefficiency. Any larger cluster size would reduce the amount of spatial reuse, without providing any extra benefit in terms of power transfer. The two extremes of this architecture are precisely the traditional multi-hop scheme, where the cluster size is 1 and the number of hops is $\sqrt{n}$, and the long-range cooperative MIMO scheme, where the cluster size is of order n and the number of hops is 1. Note also that because short-range links are bandwidth-limited and long-range links are power-limited, the network capacity is *both* bandwidth and power-limited. Thus, the capacity is sensitive to both the amount of bandwidth and the amount of power available. This regime is fundamentally a consequence of the heterogeneous nature of the links in the network.

3.5 Upper Bound on the Throughput Scaling

In this section, we derive an information-theoretic upper bound on the capacity scaling of wireless networks. The upper bound is information-theoretic because it emerges from basic assumptions on the physical channel and the network, and no specific assumption is made about the

communication or networking technique employed. As such, it characterizes the ultimate performance limit in wireless networks and applies globally to any possible network communication scheme. It turns out that the performance of the schemes described in the previous sections match asymptotically the upper bound derived below. Thus, the current and the previous sections together characterize the capacity scaling of wireless networks for the model described in Section 2.

3.5.1 Main Result

Recall the definition of the scaling exponent of the total throughput T defined earlier in Section 3.1,

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

where

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

is the scaling exponent of

$$\mathrm{SNR}_s = \frac{GP}{N_0 W(\sqrt{A/n})^\alpha}. \tag{3.17}$$

The main result of this section is to establish the following tight upper bound on the aggregate throughput scaling achieved by any scheme in the network, which matches the lower found in (3.15). The following section is devoted to the proof of this theorem.

Theorem 3.3. The scaling exponent of the aggregate throughput T is bounded above with high probability by

$$e(\alpha,\beta) \le \begin{cases} 1, & \beta \ge \alpha/2 - 1, & \text{Regime I,} \\ 2 - \alpha/2 + \beta, & \beta < \alpha/2 - 1 \text{ and } 2 \le \alpha < 3, & \text{Regime II,} \\ 1/2 + \beta, & \beta \le 0 \text{ and } \alpha \ge 3, & \text{Regime III,} \\ 1/2 + \beta/(\alpha - 2), & 0 < \beta < \alpha/2 - 1 \text{ and } \alpha \ge 3, & \text{Regime IV,} \end{cases} \tag{3.18}$$

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

The upper bounds for the dense and extended scalings extensively studied in the literature can be found as special cases of the above result. In the dense scaling, the area, the bandwidth and the power are constants that do not depend on n. It can be observed from (3.17) that $\mathrm{SNR}_s = \Theta(n^{\alpha/2})$, or equivalently dense networks correspond to $\beta = \alpha/2$, which falls in Regime I in (3.18), yielding an exponent $e(\alpha, \alpha/2) \leq 1$. This was the upper bound derived in Theorem 2.5. In the extended scaling $A = n$, while P and W are constants independent of n. In (3.17), $\mathrm{SNR}_s = \Theta(1)$ or equivalently $\beta = 0$. Thus, depending on the power path loss exponent, extended networks fall in either the second or the third regime in (3.18), with an exponent equal to

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

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

3.5.2 Cutset Upper Bound

We consider a cut dividing the network area into two equal halves. We are interested in bounding above the sum of the rates of communications passing through the cut from left to right. These communications with source nodes located on the left and destination nodes located on the right half domain are depicted in bold lines in Figure 3.1. Statistically, half of the nodes are located on the left-hand side of the network, and half of these nodes have their destination located on the right-hand side (this can be made precise using arguments similar to Lemma A.1). Hence, the above-mentioned sum-rate is equal to 1/4th of the total throughput T with high probability. The maximum achievable sum-rate between these source-destination pairs is bounded above by the capacity of the MIMO channel between all nodes S located to the left of the cut and all nodes D located to the right. Under the fast fading

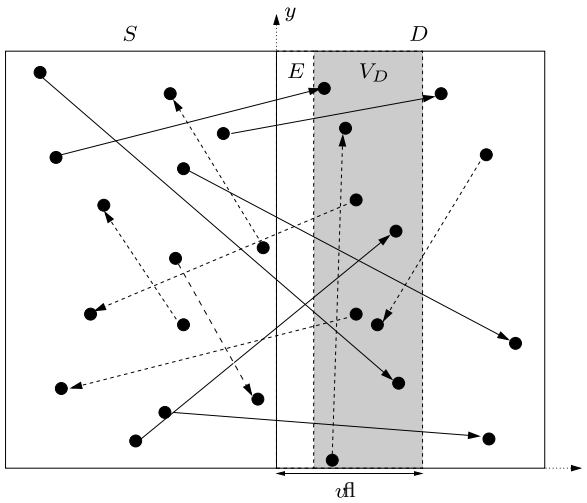

Fig. 3.1 The cut-set considered in Section 3.5.2. The communication requests that pass across the cut from left to right are depicted in bold lines.

assumption, we have (see Ref. [29] for detailed explanations)

$$T_{L\to R} \leq \max_{\substack{Q(H)\geq 0 \\ \mathbb{E}(Q_{kk}(H))\leq P/W, \forall k\in S}} \mathbb{E}\left(\log\det\left(I+\frac{1}{N_0}HQ(H)H^*\right)\right), \tag{3.19}$$

where

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

The mapping $Q(\cdot)$ is from the set of possible channel realizations H to the set of positive semi-definite transmit covariance matrices.[3] The diagonal element $Q_{kk}(H)$ corresponds to the power allocated to the kth node for channel state H. Let us simplify notation by introducing the nearest-neighbor SNR defined earlier in (2.7) and also rescale the distances in the network by this nearest-neighbor distance, defining

$$\hat{r}_{ik}:=\frac{1}{\sqrt{A/n}}r_{ik} \quad \text{and} \quad \hat{H}_{ik}:=\frac{e^{j\theta_{ik}}}{\hat{r}_{ik}^{\alpha/2}}. \tag{3.20}$$

[3] Indeed, full channel state information is assumed at the transmitter nodes; hence, the transmit covariance matrix can be tuned according to the channel state realizations.

Note that the first transformation rescales space and maps our original network of area $\sqrt{A} \times \sqrt{A}$ to a network of area $\sqrt{n} \times \sqrt{n}$. Consequently, the matrix $\hat{H}$ defined in terms of the rescaled distances relates to such a network with area n. Normalizing the typical nearest-neighbor distance to 1 provides the convenience that the received SNR in a point-to-point transmission between two nodes at rescaled distance $\hat{r}$ can be simply expressed as $\mathrm{SNR}_s \hat{r}^{-\alpha}$. We can thus rewrite (3.19) in terms of these new variables as[4]

$$T_{L\to R} \leq \max_{\substack{Q(\hat{H})\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}))\leq 1, \forall k\in S}} \mathbb{E}\left(\log\det(I + \mathrm{SNR}_s \hat{H}Q(\hat{H})\hat{H}^*)\right). \quad (3.21)$$

One way to upper bound (3.21) is through upper bounding the capacity by the total received SNR, formally using the relation

$$\log\det(I + \mathrm{SNR}_s \hat{H}Q(\hat{H})\hat{H}^*) \leq \mathrm{Tr}(\mathrm{SNR}_s \hat{H}Q(\hat{H})\hat{H}^*). \quad (3.22)$$

The upper bound is tight only if the SNR received by each right-hand side node, i.e., each diagonal entry of the matrix $\mathrm{SNR}_s \hat{H}Q(\hat{H})\hat{H}^*$, is small. (Note that the relation in (3.22) relies on the inequality $\log(1+x) \leq x$, which is only tight if x is small.) Whether this is the case or not depends on SNR_s. It can be shown that if $\mathrm{SNR}_s \leq 1$, the network is highly power-limited and the received SNR is small, i.e., decays to zero with increasing n, for every right-hand side node. Using (3.22) will yield a tight upper bound in that case. However, in the general case, SNR_s can be arbitrarily large, which can result in high received SNR for certain right-hand side nodes that are located close to the cut or even for all nodes in D. Hence, before using (3.22), we need to distinguish between those nodes in D that receive high SNR and those that have poor power connections to the left-hand side.

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

Assumption 3.1. For the sake of simplicity in the presentation, we assume in this section that there is a rectangular region located immediately to the right of the cut that is cleared of nodes. Formally, we assume that the set of nodes $E = \{i \in D : 0 \leq \hat{x}_i \leq 1\}$ is empty, where $\hat{x}_i$ denotes the horizontal coordinate of the rescaled position $\hat{r}_i = (\hat{x}_i, \hat{y}_i)$ of node i. In fact, w.h.p., this property does not hold in a random realization of the network. However, making this assumption allows to exhibit the central ideas in the following derivation in a simpler manner. A rigorous derivation of the result (without this particular assumption) can be found in Ref. [22].

Let V_D denote the set of nodes located on a rectangular strip immediately to the right of the empty region E. Formally, $V_D = \{i \in D : 1 \leq \hat{x}_i \leq \hat{v}\}$, where $1 \leq \hat{v} \leq \sqrt{n}/2$ and $\hat{v} - 1$ is the rescaled width of the rectangular strip V_D (Figure 3.1). We would like to tune $\hat{v}$ so that V_D contains the right-hand side nodes with high received SNR from the left-hand side; i.e., those nodes whose received SNR is larger than a threshold, say 1. Note, however, that we do not yet know the covariance matrix Q of the transmissions from the left-hand side nodes, which is to be determined from the maximization problem in Equation (3.21). Thus, we cannot really compute the received SNR of a right-hand side node. For the purpose of specifying V_D, however, let us arbitrarily look at the case where Q is the identity matrix and let us define the received SNR of a right-hand side node $i \in D$, when left-hand side nodes are transmitting *independent* signals at full power, to be

$$\mathrm{SNR}_i := \sum_{k \in S} \frac{P}{N_0 W} |H_{ik}|^2 = \mathrm{SNR}_s \sum_{k \in S} |\hat{H}_{ik}|^2 = \mathrm{SNR}_s \hat{d}_i, \tag{3.23}$$

where we have defined

$$\hat{d}_i := \sum_{k \in S} |\hat{H}_{ik}|^2. \tag{3.24}$$

Later, we will see that this arbitrary choice of identity covariance matrix is indeed asymptotically optimal (Lemma 3.5). A good approximation

for $\hat{d}_i$ is

$$\hat{d}_i \approx \hat{x}_i^{2-\alpha}, \tag{3.25}$$

where $\hat{x}_i$ denotes the rescaled horizontal coordinate of the right-hand side node $i \in D$. (This fact is made precise in Lemma 3.7.) Recall that $1 \le \hat{x}_i \le \sqrt{n}/2$ and as $\alpha \ge 2$, $\hat{d}_i$ is decreasing in $\hat{x}_i$. Using (3.23) and (3.25), we can identify three different regimes and specify $\hat{v}$ accordingly:

(1) If $\mathrm{SNR}_s \ge n^{\alpha/2-1}$, then $\mathrm{SNR}_i \gtrsim 1, \forall i \in D$. Hence, let us choose $\hat{v} = \sqrt{n}/2$ or equivalently $V_D = D$ in this case.
(2) If $\mathrm{SNR}_s \le 1$, then $\mathrm{SNR}_i \lesssim 1, \forall i \in D$. Thus, let us choose $\hat{v} = 1$ or equivalently $V_D = \emptyset$.[5]
(3) If $1 < \mathrm{SNR}_s < n^{\alpha/2-1}$, then let us choose

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

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

We now would like to break the information transfer from the left-half domain S to the right-half domain D in (3.21) into two terms. The first term governs the information transfer from S to V_D. The second term governs the information transfer from S to the remaining nodes on the right-half domain, i.e., $D \backslash V_D$. Recall that the characteristic of the nodes in V_D is that they have good power connections to the left-hand side, that is, the information transfer from S to V_D is not limited in power, but can be limited in degrees of freedom. Thus, it is reasonable to bound the rate of this first information transfer by the cardinality of the set V_D, rather than the total received SNR. On the other hand, the remaining nodes in $D \backslash V_D$ have poor power connections to the left-half domain and the information transfer to these nodes is limited in power; hence, using (3.22) is tight. Formally, we proceed by applying the generalized block Hadamard's inequality (also known as Fischer's inequality), which yields

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

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

where $\hat{H}_1$ and $\hat{H}_2$ are obtained by partitioning the original matrix $\hat{H}$: $\hat{H}_1$ is the rectangular matrix with entries $\hat{H}_{ik}, k \in S, i \in V_D$ and $\hat{H}_2$ is the rectangular matrix with entries $\hat{H}_{ik}, k \in S, i \in D \backslash V_D$. In turn, Equation (3.21) is bounded above by

$$\begin{aligned} T_{L\to R} \leq & \max_{\substack{Q(\hat{H}_1)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_1))\leq 1, \forall k\in S}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \hat{H}_1 Q(\hat{H}_1)\hat{H}_1^*)) \\ & + \max_{\substack{Q(\hat{H}_2)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_2))\leq 1, \forall k\in S}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \hat{H}_2 Q(\hat{H}_2)\hat{H}_2^*)). \end{aligned} \tag{3.26}$$

The first term in (3.26) can be bounded by applying Hadamard's inequality once more, or equivalently, by considering the sum of the capacities of the individual multiple-input single-output (MISO) channels between nodes in S and each node in V_D,

$$\begin{aligned} & \max_{\substack{Q(\hat{H}_1)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_1))\leq 1, \forall k\in S}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \hat{H}_1 Q(\hat{H}_1)\hat{H}_1^*)) \\ & \leq \sum_{i\in V_D} \log\left(1 + n\,\mathrm{SNR}_s \sum_{k\in S} |\hat{H}_{ik}|^2\right), \qquad (3.27) \\ & \leq (\hat{v}-1)\sqrt{n}\log n \log(1 + n^{2+\alpha(1/2+\delta)}\mathrm{SNR}_s), \qquad (3.28) \end{aligned}$$

w.h.p. for any $\delta > 0$. Inequality (3.27) comes from the fact that for any covariance matrix Q of the transmissions from S, the SNR received by each node $i \in V_D$ is smaller than $n\,\mathrm{SNR}_s \hat{d}_i$. Inequality (3.28) is obtained by using the crude bound $\hat{d}_i \leq n^{1+\alpha(1/2+\delta)}$, which follows from the fact that the rescaled minimal separation between any two nodes in the network is larger than $1/n^{1/2+\delta}$ w.h.p. for any $\delta > 0$ (Lemma A.1(a)) and the number of nodes in S are smaller than n. On the other hand, the number of nodes in V_D is upper bounded by $(\hat{v}-1)\sqrt{n}\log n$ w.h.p (Lemma A.1(b)).

The second term in (3.26) is the capacity of the MIMO channel between nodes in S and nodes in $D \backslash V_D$. The following lemma provides an upper bound on the capacity of this channel. Although the main idea is to upper bound the capacity by the total received SNR using

inequality (3.22), this is not done immediately as we first need to waive out the possibility of communicating only through non-typically good channel matrices. Once inequality (3.22) is applied, we need to handle the maximization over all admissible covariance matrices that are allowed to be functions of the channel state realizations.

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

Lemma 3.5. Let $\mathrm{SNR}_{\mathrm{tot}}$ be the total SNR received by all the nodes in $D\backslash V_D$, when nodes in S are transmitting *independent* signals at full power, i.e.,

$$\mathrm{SNR}_{\mathrm{tot}} := \sum_{i\in D\backslash V_D} \mathrm{SNR}_i = \mathrm{SNR}_s \sum_{i\in D\backslash V_D} \hat{d}_i. \tag{3.29}$$

Recall that SNR_i has been defined in (3.23) as the SNR received by the node $i \in D$ under independent transmissions from the left-hand side. Then, for every $\epsilon > 0$,

$$\max_{\substack{Q(\hat{H}_2)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_2))\leq 1, \forall k\in S}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \hat{H}_2 Q(\hat{H}_2)\hat{H}_2^*)) \leq n^{\epsilon}\mathrm{SNR}_{\mathrm{tot}}. \tag{3.30}$$

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

$$\mathrm{SNR}_{\mathrm{tot}} \leq \begin{cases} K_1 \mathrm{SNR}_s n^{2-\alpha/2}(\log n)^3, & 2 \leq \alpha < 3, \\ K_1 \mathrm{SNR}_s \hat{v}^{3-\alpha}\sqrt{n}(\log n)^3, & \alpha \geq 3, \end{cases} \tag{3.31}$$

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

Lemma 3.5 says couple of surprising things. First of all, it says that independent signaling at the transmit nodes is sufficient to achieve the cutset upper bound (up to a multiplicative factor of order n^{ε}). Note that, *a priori*, on the left-hand side of (3.30), nodes are allowed to cooperate and do any sort of transmit beamforming over channel state realizations. Lemma 3.5 says that this is not necessary. This explains

why we earlier based our choice of $\hat{v}$ on the assumption of independent transmissions from the left-hand side nodes. Independent signaling is indeed good enough, at least as far as the scaling of the capacity is concerned.

Second, depending on α, the lemma identifies a dichotomy on how the received SNR under independent transmissions scales with system size (3.31). This dichotomy can be interpreted as follows. The total received SNR is dominated either by the power transferred between node pairs separated by a relatively short distance (of the order of $\hat{v}$) or by the power transferred between nodes far away (at distance of the order of $\sqrt{n}$). There are relatively fewer node *pairs* at distance $\hat{v}$; however, the channels between these pairs are considerably stronger than the pairs at distance $\sqrt{n}$. When the attenuation parameter α is less than 3, the received power is dominated by the transfer between the large number of node pairs at distance $\sqrt{n}$. There are n^2 node pairs separated by a rescaled distance of the order of $\sqrt{n}$, which yields a total SNR transfer of $\mathrm{SNR}_s \times n^2 \times \sqrt{n}^{-\alpha}$ between these pairs. This is the first term in (3.31), up to logarithmic terms. When $\alpha \geq 3$, the received SNR in the cutset bound is dominated by the power transferred between node pairs at distance $\hat{v}$. There are an order of $\sqrt{n} \times \hat{v}^3$ pairs located at distance of the order of $\hat{v}$. (Consider the nodes in S located up to $\hat{v}$ rescaled horizontal distance to the cut and those nodes in $D \backslash V_D$ located up to $2\hat{v}$ horizontal distance to the cut. Then count the number of node pairs that are separated with a distance of the order of $\hat{v}$.) Hence, the total SNR transfer between these node pairs is equal to $\sqrt{n}\hat{v}^3 \times (\hat{v})^{-\alpha}$. This argument yields the second term in (3.31), up to logarithmic terms.

Combining the upper bounds (3.28) and (3.30) together with our choices for $\hat{v}$ specified earlier, we get an upper bound on $T_{L \to R}$ in terms of SNR_s and n. Here, we state the final result in terms of scaling exponents. We have

$$e(\alpha,\beta) \leq \begin{cases} 1, & \beta \geq \alpha/2 - 1, \\ 2 - \alpha/2 + \beta, & \beta < \alpha/2 - 1 \quad \text{and} \quad 2 \leq \alpha < 3, \\ 1/2 + \beta, & \beta \leq 0 \quad \text{and} \quad \alpha \geq 3, \\ 1/2 + \beta/(\alpha - 2), & 0 < \beta < \alpha/2 - 1 \quad \text{and} \quad \alpha \geq 3, \end{cases} \tag{3.32}$$

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

Note that in the first regime, Equation (3.28) gives the dominant term for the upper bound on the capacity with $\hat{v} = \sqrt{n}/2$ (or equivalently $V_D = D$), while (3.30) is zero. The capacity of the network is limited by the degrees of freedom of an $n \times n$ MIMO transmission between the left-hand side and the right-hand side nodes.

In the second regime, $\hat{v} = \sqrt{n}/2$ also and Equation (3.30) together with the corresponding upper bound being the first line of (3.31) yields a larger contribution than (3.28). The capacity is limited by the total received SNR in a MIMO transmission between the left-hand side nodes and $D \backslash V_D$. Note that this total received SNR is equal (in order) to the power transferred in a MIMO transmission between two groups of n nodes separated by a distance of the order of the diameter of the network, i.e., $n^2 \times (\sqrt{n})^{-\alpha} \times \mathrm{SNR}_s$.

In the third regime, the capacity is determined by (3.30) with $\hat{v} = 1$, or equivalently $V_D = \emptyset$; hence, (3.28) is zero. The corresponding upper bound is the second line of (3.31). The capacity in this regime is still limited by the total SNR received by nodes in $D \backslash V_D$ ($= D$ now); however, the total is now dominated by the SNR transferred between the nearest nodes to the cut, i.e., $\sqrt{n}$ pairs separated by the nearest-neighbor distance ($\hat{v} = 1$), yielding $\sqrt{n} \times \mathrm{SNR}_s$. Note that this is where we make use of the assumption that there are no nodes located at rescaled distance smaller than 1 to the cut. Owing to this assumption, the choice $\hat{v} = 1$ vanishes the upper bound (3.28) and simultaneously yields $K_1 \mathrm{SNR}_s \sqrt{n} (\log n)^2$ in the last line in (3.31) for the total SNR transferred from S to D. If there were nodes closer than rescaled distance 1 to the cut, we would need to choose $\hat{v} < 1$ to vanish the contribution from (3.28), which would yield a larger value for the term $K_1 \mathrm{SNR}_s \hat{v}^{3-\alpha} \sqrt{n} (\log n)^2$, as $\alpha \geq 3$ in this regime. The difficulty is the following. We would like to conclude that in this regime, the power transfer between left- and right-hand side nodes is dominated by the contribution of the order $\sqrt{n}$ nearest-neighbor pairs located around the cut. However, there can be a pair of nodes, one node on the left and the other one on the right of the cut, which is separated by a distance much smaller than the nearest-neighbor distance in the network

and the capacity of the channel between these two nodes can be much larger than the total contribution of the $\sqrt{n}$ nearest-neighbor pairs. Even though this may be the case for the cut considered, it is not possible to rely on such pairs for communicating inside the network, as these pairs do not form a path inside the network w.h.p. This fact is made precise in Ref. [22].

The most interesting regime is the fourth one. Both (3.28) and (3.30), with the choice $\hat{v} = \mathrm{SNR}_s^{1/(\alpha-2)}$, yield the same contribution. Note that (3.28) upper bounds the information transfer to V_D, the set of nodes that have bandwidth-limited connections to the left-hand side. This information transfer is limited in degrees of freedom. On the other hand, (3.30) upper bounds the information transfer to $D \backslash V_D$, the set of nodes that have power-limited connections to the left-hand side. This second information transfer is power-limited. Therefore, in this regime, the network capacity is limited in both degrees of freedom and power, as increasing the bandwidth increases the first term (3.28) and increasing the power increases the second term (3.30). This behavior is a consequence of the heterogeneous nature of links in a network and does not occur in point-to-point links. □

Proof of Lemma 3.5. We are interested in the scaling of the MIMO capacity,

$$\max_{\substack{Q(\hat{H}_2)\geq 0 \\ \mathbb{E}(Q_{kk}(\hat{H}_2))\leq 1, \forall k \in S}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \hat{H}_2 Q(\hat{H}_2)\hat{H}_2^*)). \tag{3.33}$$

A natural way to upper bound (3.33) is to first relax the individual power constraint

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

to a total power constraint

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

where $|S|$ denotes the cardinality of the set S. In the present context, however, this is not convenient, as the matrix $\hat{H}_2$ is badly conditioned: some nodes in S are close to the cut and some are far apart; hence, the impact of their 1 Watt power on the system performance is quite

different. A total transmit power constraint allows the transfer of power from the nodes far away from the cut to those nodes that are located close to the cut, resulting in a loose bound. Instead, we will relax the individual power constraints to a total *weighted* power constraint, where the weight assigned to a node is proportional to the impact of its unit power. The impact is measured by the total *received* power on the right-hand side of the cut, per Watt of transmit power from that left-hand side node.

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

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

We define the normalized matrix

$$\tilde{H}_{ik} = \frac{1}{\sqrt{w_k}} \hat{H}_{ik}, \quad i \in D \backslash V_D, \;\; k \in S. \tag{3.34}$$

The expression (3.33) is then equal to

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

Note that $\mathrm{SNR}_s w_k$ corresponds to the total received SNR by the nodes in $D \backslash V_D$ of the signal sent by the user $k \in S$. Having weighted each of the individual power constraints in (3.33) by their impact, we now relax them to a total power constraint, which yields the following upper bound for (3.33),

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

where

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

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

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

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

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

for all n. The following more precise statement is shown in Ref. [22].

Lemma 3.6. There exists $K_2 > 0$ such that for all $l \geq 0$ and all n,

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

The method of proof relies on the analysis done for matrices with i.i.d. entries (see, for example, Ref. [2]). The main difference in the present case is that entries do not share all the same variance, which explains the appearance of an additional $\log n$ factor in the upper bound on the largest singular value of $\tilde{H}$.

As a consequence of this lemma, we obtain, using Chebychev's inequality,

$$\mathbb{P}(B_{n,\varepsilon}) = \mathbb{P}(\|\tilde{H}\|^2 > n^{\varepsilon}) \leq \frac{\mathbb{E}(\|\tilde{H}\|^{2l})}{n^{\varepsilon l}} \leq K_2\, n^{1-\varepsilon l} (\log n)^3 \leq K_2\, n^{-p}, \tag{3.35}$$

where $p \geq 1$ can be chosen arbitrarily large by choosing l accordingly.

It follows that

$$\begin{aligned}
&\max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{\mathrm{tot}}}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)) \\
&\quad\leq \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{\mathrm{tot}}}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)1_{B_{n,\varepsilon}}) \\
&\quad\quad + \max_{\substack{\tilde{Q}(\tilde{H})\geq 0 \\ \mathbb{E}(\mathrm{Tr}\tilde{Q}(\tilde{H}))\leq W_{\mathrm{tot}}}} \mathbb{E}(\log\det(I + \mathrm{SNR}_s \tilde{H}\tilde{Q}(\tilde{H})\tilde{H}^*)1_{B_{n,\varepsilon}^c}).
\end{aligned} \tag{3.36}$$

The first term in (3.36) refers to the event that the channel matrix $\tilde{H}$ is accidentally ill-conditioned. As the probability of such an event

is polynomially small by (3.35), the contribution of this first term is actually negligible. In the second term in (3.36), the matrix $\tilde{H}$ is well-conditioned, and this term is actually proportional to the maximum SNR transfer from S to $D\backslash V_D$. Details follow below.

For the first term in (3.36), we use Hadamard's inequality and obtain

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

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

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

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

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

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

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

which follows from the definition of $\tilde{H}$ in (3.34). The fact that the rescaled minimum distance between the nodes in S and $D\backslash V_D$ is at

least 1 yields

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

Note that $x \mapsto x\log(1+1/x)$ is increasing on $[0,1]$ and using (3.35), we finally obtain that for any $p \geq 1$, there exists $K_2 > 0$ such that

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

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

For the second term in (3.36), we simply have

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

The last thing that therefore needs to be checked is the scaling of W_{tot} stated in Lemma 3.5.

Let us divide the rescaled network area of size n into n squarelets of area 1. By Part (b) of Lemma A.1, there are no more than $\log n$ nodes in each squarelet, with high probability. Let us consider grouping the squarelets on the left of the cut into $\sqrt{n}$ horizontal rectangular areas of height 1 and width $\sqrt{n}/2$, as shown in Figure 3.2. Let S_m denote the nodes in S that are located on the mth rectangle, so that $S = \bigcup_{m=1}^{\sqrt{n}} S_m$. We are interested in bounding above

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

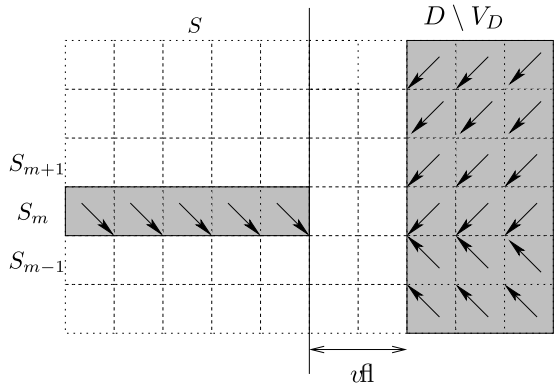

Fig. 3.2 The displacement of the nodes inside the squarelets to squarelet vertices, indicated by arrows.

Let us consider, for a given m,

$$\sum_{k \in S_m} w_k = \sum_{k \in S_m, i \in D \setminus V_D} |\hat{H}_{ik}|^2 = \sum_{k \in S_m, i \in D \setminus V_D} \hat{r}_{ik}^{-\alpha}. \tag{3.37}$$

Note that if we move the points that lie in each squarelet of S_m together with the nodes in the squarelets of $D \setminus V_D$ onto the squarelet vertex as indicated by the arrows in Figure 3.2, all the terms in the summation in (3.37) can only increase, as the displacement can only decrease the Euclidean distance between the nodes involved. Note that the modification results in a regular network with at most $\log n$ nodes at each squarelet vertex on the left and at most $2 \log n$ nodes at each squarelet vertex on the right. Considering the same reasoning for all rectangular slabs $S_m, m = 1, \ldots, \sqrt{n}$ allows to conclude that W_{tot} for the random network is with high probability less than the same quantity computed for a regular network where nodes are located on a square grid of distance 1, with $\log n$ nodes at each left-hand side vertex and $2 \log n$ nodes at each right-hand side vertex.

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

and $i_y = 0, \ldots, \sqrt{n}$, so that

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

and

$$w_{k_x,k_y} = \sum_{i_x=\hat{v}}^{\sqrt{n}/2} \sum_{i_y=0}^{\sqrt{n}} \frac{1}{((i_x + k_x)^2 + (i_y - k_y)^2)^{\alpha/2}}, \tag{3.38}$$

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

$$W_{\text{tot}} \leq 2(\log n)^2 \sum_{k_x=0}^{\sqrt{n}/2} \sum_{k_y=0}^{\sqrt{n}} w_{k_x,k_y}. \tag{3.39}$$

The following lemma establishes the scaling of w_{k_x,k_y} defined in Equation (3.38).

Lemma 3.7. There exist constants $K_3, K_4 > 0$ independent of k_x, k_y and n such that

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

and

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

The rigorous proof of the lemma is given in Ref. [22]. A heuristic way of thinking about the approximation

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

can be obtained through the Laplace principle. The summation in w_{k_x,k_y} scales the same as the maximum term in the sum times the number of terms, which have roughly this maximum value. The maximum term is of the order of $1/(\hat{v} + k_x)^{\alpha}$. The terms that take on roughly this value are those for which i_x runs from $\hat{v}$ to the order of

$2\hat{v} + k_x$ and i_y runs from k_y to k_y plus or minus the order of $\hat{v} + k_x$. There are roughly $(\hat{v} + k_x)^2$ such terms. Hence, $w_{k_x,k_y} \approx 1/(\hat{v} + k_x)^\alpha \cdot (\hat{v} + k_x)^2 = (\hat{v} + k_x)^{2-\alpha}$.

We can now use the upper bound given in Lemma 3.7, which gives

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

for another constant $K_5 > 0$ independent of n. This upper bound combined with (3.39) yields (3.31) and completes the proof of Lemma 3.5. □

4

Space

The results presented in the two previous sections crucially rely on the fact that the degrees of freedom of MIMO transmissions are of the order of the number of nodes participating to the transmission. This fact was established in Equation (2.19) at the end of Section 2, based on the assumption that the phase shifts $\phi_{ik}[m]$ in the channel model (2.2) are uniformly distributed on $[0, 2\pi]$ and independent across node pairs. This is a standard model in wireless communication based on a far-field assumption: under free-space propagation, the phase shifts derived from Maxwell's equations are given by

$$\phi_{ik} = 2\pi r_{ik}/\lambda,$$

where r_{ik} is the distance between nodes i and k, and λ is the carrier wavelength. When the distances between the nodes are at a much larger spatial scale compared with λ, so that the phases ϕ_{ik} get completely mixed even when nodes move insignificantly, the phases are modeled as completely random and independent of the actual positions. However, in a large network regime there is a large number of phase variables, n^2 of them. It is not clear how large the spatial separation between the nodes, or equivalently the area of the network, should be for such

a large collection of variables to behave approximately independent of each other.

The goal of this section is to develop a quantitative basis for the intuitive justification of the i.i.d. phase model: we identify the conditions under which the i.i.d. phase model hence the conclusions of the earlier two sections hold in wireless networks. We also investigate regimes where the i.i.d. phase model fails to hold. Such regimes are space-limited. We extend the approximate characterization of the capacity and our discussion on optimal architectures from the previous sections to space-limited regimes. We show that the distributed MIMO-based architectures developed in the earlier sections provide significant capacity gains over multihop even when the i.i.d. phase model fails to hold.

4.1 Model

As in the previous sections, we assume that nodes are randomly spread out in a square planar region of area A. We assume that each node has a device equipped with one antenna, oriented in the direction perpendicular to the plane. Using a standard dipole model for the antennas and assuming free-space propagation, one can derive the following expression from Maxwell's equations for the channel attenuation coefficient between nodes i and k:

$$H_{ik} = \sqrt{G}\frac{\exp(2\pi j r_{ik}/\lambda)}{r_{ik}}, \tag{4.1}$$

where λ is the carrier wavelength, r_{ik} is the distance between nodes i and k and G is the Friis constant given by

$$G = G_t G_r \left(\frac{\lambda}{4\pi}\right)^2$$

with G_t and G_r being the transmit and receive antenna gains, respectively. The goal of the current section is to rethink some of the conclusions of the earlier sections based on this more fundamental physical channel model, which is a direct consequence of Maxwell's equations for free-space propagation. In particular, we characterize the scaling exponent

$$e_{\text{phy}}(\nu) := \lim_{n\to\infty} \frac{\log T(n,\nu)}{\log n}$$

of the best achievable aggregate throughput $T(n) = nR(n)$ under this new channel model for any $\nu > 0$ when $A = n^{\nu}$. In complete analogy with the power discussion of Section 3, this characterization provides an understanding of the impact of the area of the network, the space, on the capacity of wireless networks and of the space-limited operating regimes of these networks. We do not address in this section the *joint* effect of both power and space limitations. This would require the characterization of the scaling exponent under this new channel model for the complete interplay between the system parameters A, P and W. This is an open problem. Instead, we specify later conditions to avoid power limitation in the network.

Let us point towards the main differences between this model and the model (2.2) considered in Sections 2 and 3. The phase shifts in (4.1) are deterministic functions of the distances between the nodes, as opposed to being independent and uniformly distributed random variables in (2.2); consequently, they do not vary over time, contrary to what was assumed in (2.2). Finally, the path loss exponent α is equal to 2 in this model, because of the free-space propagation assumption.

Considering the above line-of-sight model might seem restrictive *a priori*, as it is well known that in practical scenarios, there are additional fading effects, mainly due to scattering and multi-path propagation. Adding such effects into the picture would actually bring us back to the first model (2.2), under which the efficiency of MIMO transmissions has already been demonstrated. Instead, the line-of-sight model corresponds to the worst-case scenario, where the MIMO transmissions are the least efficient and spatial limitation is therefore expected to be the most severe. We will see nevertheless in the following that distributed MIMO transmissions can still improve the aggregate throughput scaling of wireless networks under this free-space model.

4.2 Upper Bound on the Throughput Scaling

The work [7] of Franceschetti et al. was the first to show that the predictions of the i.i.d. phase model in (2.2) and the physical channel model introduced in this section can be different; the capacity of wireless networks can suffer from space limitation, which is not predicted by

the random phase model. In this section, we review the result obtained in Ref. [7]. The main assumptions made in there are the following:

(1) The basic channel model is a variant of the line-of-sight model (4.1) derived from Maxwell's equations, which exhibits the same features as the one presented here. In addition, Ref. [7] also considers the presence of scatterers.
(2) The area A of the network is assumed to grow linearly with the number of nodes n, so as to ensure a constant density of nodes.

Under these assumptions, the following upper bound is obtained in Ref. [7] on the throughput scaling of the network:

$$T(n) = O(\sqrt{n}(\log n)^2). \tag{4.2}$$

Let us give here a glimpse of the key ideas behind this result, as described in Ref. [7]:

- The result is obtained via the upper bounding technique discussed in Section 3.5.2: the network throughput is upper bounded by the maximum information transfer between two subsets of nodes of equal size. A thin layer of constant width is assumed to separate these two subsets.
- The discrete problem is first translated into a continuous one; the $n \times n$ channel matrix H between the two subsets of nodes therefore becomes a continuous operator describing the propagation of the electromagnetic field over space.
- The propagation operator is then decomposed into three parts and a refined analysis of the spectral properties of the operator over the above-mentioned separating layer shows that it can only convey up to $O(\sqrt{n}\log n)$ independent signals in this region, limiting therefore the total capacity of the transmission by this amount (the extra $\log n$ factor in the result comes from the potential power gain due to the presence of n transmitting nodes).

The above result can at first lead to the conclusion that the total degrees of freedom in the network is upper bounded by $\sqrt{n}$ due to the constraints imposed by the physical channel and that multi-hop

is scaling optimal. A deeper look from the perspective of the multi-parameter formulation of Section 3.1 reveals that this not quite the case. If we look at the more general scaling law problem of characterizing the scaling of the capacity when

$$A = n^{\nu} \quad \text{for any } \nu > 0,$$

under the new channel model, we can uncover the individual dependencies of the capacity to A and n. In this case, the upper bound (4.2) translates into[1]

$$T(n) = \begin{cases} O(\sqrt{n}(\log n)^2), & \text{if } \sqrt{A}/\lambda \ll \sqrt{n}, \\ O(\sqrt{A}/\lambda(\log(\sqrt{A}/\lambda)^2), & \text{if } \sqrt{n} \ll \sqrt{A}/\lambda \ll n, \\ O(n(\log n)^2), & \text{if } \sqrt{A}/\lambda \gg n. \end{cases} \tag{4.3}$$

The limitation uncovered in Ref. [7] via the physical channel model is therefore not a universal limitation depending only on the number of nodes; it is rather about space. The second line of (4.3) says that the throughput of the network is limited by the *diameter* of the network (normalized by the carrier wavelength), a quantity that can be interpreted as the number of spatial degrees of freedom available in the network. The throughput cannot scale faster than this number. The conclusion that the capacity cannot scale faster than $\sqrt{n}$ in (4.2) comes from the assumption that A grows linearly in n, so that $\sqrt{A}/\lambda$ is proportional to $\sqrt{n}$. When $\sqrt{A}/\lambda$ is as small as $\sqrt{n}$, the first line of (4.3) says that the throughput cannot scale faster than $\sqrt{n}$ (up to logarithmic terms). In this case, the spatial degrees of freedom available in the network are so few that they can be already exploited by multi-hop. (It is easy to verify that the multi-hop architecture discussed in Section 2.2 achieves the same $\sqrt{n}$ scaling under the new channel model.) More sophisticated cooperation is useless in this case.

[1] As usual, this result is based on characterizing the scaling exponent of the aggregate throughput. In terms of the scaling exponent, we have

$$e_{\text{phy}}(\nu) \leq \begin{cases} 1/2, & \text{if } \nu < 1, \\ \nu/2, & \text{if } 1 \leq \nu < 2, \\ 1, & \text{if } \nu \geq 2. \end{cases}$$

However, as it is more informative, in this section we skip the statement of the results in terms of the scaling exponent and only refer to the corresponding approximations of the capacity as in (4.3).

However, for actual networks, there is a huge difference between $\sqrt{A}/\lambda$ and $\sqrt{n}$ as illustrated by the numerical example below. In such a case, (4.3) allows for much better scaling than $\sqrt{n}$, a scaling linear in $\min(\sqrt{A}/\lambda, n)$ up to logarithmic terms. In particular, when $\sqrt{A}/\lambda$ is larger than n it allows for linear scaling in n, while the performance of multi-hop is always $\sqrt{n}$ independent of $\sqrt{A}/\lambda$. The remaining question we investigate in the following section is whether these additional degrees of freedom can be exploited by the distributed MIMO-based architectures developed in the previous sections. Earlier, we have investigated the benefits of these architectures based on the i.i.d. phase model. Here, by investigating their performance under the more fundamental physical channel model, we also gain insight about the regimes where conclusions from the i.i.d. phase model hold or fail to hold.

Example 4.1. Take an example of a network serving $n = 10'000$ students on a campus of $1\,\mathrm{km}^2$, operating at $3\,\mathrm{GHz}$: the corresponding carrier wavelength is $0.1\,\mathrm{m}$. $\sqrt{A}/\lambda = 10'000$, while $\sqrt{n}$ is only 100, two orders of magnitude smaller. In such a network, the spatial degrees of freedom available for communication are many more than what can be exploited by multi-hop. Indeed here, $\sqrt{A}/\lambda$ is comparable with n. Even though 10'000 users on $1\,\mathrm{km}^2$ seems like a pretty dense network, there are still sufficient spatial degrees of freedom for all the users in the system; we are at the boundary between the second and third regimes in Equation (4.3).

4.3 Optimal Cooperation

The main result of this section can be summarized as follows: provided there is enough available power (this is to be made precise below), the upper bounds on the throughput scaling obtained in Equation (4.3) are all achievable, up to logarithmic factors. Accordingly, the optimal operation of the network falls into three different operating regimes:

(1) $\sqrt{A}/\lambda \ll \sqrt{n}$: The number of spatial degrees of freedom is too small, more sophisticated cooperation is useless and nearest-neighbor multi-hopping is optimal.

(2) $\sqrt{A}/\lambda \gg n$: The number of spatial degrees of freedom is n, the optimal performance can be achieved by the same hierarchical cooperation scheme introduced in Section 2 and provides dramatic gain over multi-hop. Spatial degree of freedom limitation does not come into play and the performance is *as though* phases were i.i.d. uniform across node pairs.

(3) $\sqrt{n} \ll \sqrt{A}/\lambda \ll n$: The number of degrees of freedom is smaller than n; hence, the spatial limitation is felt, but larger than what can be achieved by simple multi-hopping. A modification of the hierarchical cooperation scheme achieves the optimal scaling in this regime.

Note that it is the second case, $\sqrt{A}/\lambda \gg n$, when the physical channel model yields the same conclusion with the i.i.d. phase model from the previous sections: provided that there is sufficient power hierarchical cooperation can achieve linear scaling. This identifies $\sqrt{A}/\lambda \gg n$ as the regime where the i.i.d. phase model is appropriate. The other two regimes for $\sqrt{A}/\lambda \ll n$ are newly uncovered by the physical channel model and are not predicted by the i.i.d. phase model.

The precise statement of the result is given in the following theorem.

Theorem 4.2. Consider a wireless network of n nodes distributed uniformly at random over a square area A such that $A = n^\nu$ for $\nu \geq 1$ and assume that the long-distance SNR in this network defined in (3.10) is greater than or equal to 0 dB. Then for any $\varepsilon > 0$, there exists a constant $K > 0$ independent of n and A such that the following throughput can be achieved with high probability as n grows large:

$$T(n) \geq K \min(n, \sqrt{A}/\lambda)^{1-\varepsilon}.$$

Recall from Section 3 that the condition $\mathrm{SNR}_l \geq 0\,\mathrm{dB}$ ensures that the network is not power-limited; there is enough power available for the long-range MIMO transmissions, at all levels of the hierarchy. Note that as the path loss attenuation α is equal to 2 in the present case, the

long-range SNR of the network is also equal to the short-range SNR (see Equation (3.9)).

The key for the proof of the above theorem is to characterize the degrees of freedom of the MIMO transmissions used in the hierarchical cooperation scheme under the new physical channel model. In the following sections, we prove Theorem 4.2 by capitalizing on a result on the spatial degrees of freedom in MIMO, which is later proved in Section 4.4.

4.3.1 Spatial Degrees of Freedom of MIMO Transmissions

It has been shown at the end of Section 2 that under the random phase model (2.2), the number of bits that can be transmitted simultaneously in a MIMO transmission between two clusters of M nodes is of order M (provided a high enough SNR between the two clusters). This is due to the independence assumption regarding the random phase shifts $\phi_{ik}[m]$ in (2.2), which implies that the channel matrix H between the two clusters is full rank with high probability. Under the line-of-sight model (4.1), the phase shifts are linear functions of the inter-node distances and are therefore strongly correlated; hence, it is not guaranteed anymore that the matrix H is full rank.

As we show in Section 4.4, it turns out that the number of spatial degrees of freedom in a MIMO transmission is still relatively high under the line-of-sight model, provided the area occupied by each cluster is large and the distance between the two clusters is not too large compared with the diameter of the clusters. More precisely, it can be shown that under the line-of-sight model (4.1), the capacity of a MIMO transmission between two clusters of M nodes is linear in M as soon as (up to logarithmic factors)

$$\frac{A_c}{\lambda d} \geq M, \tag{4.4}$$

provided the long-distance SNR between these two clusters defined as

$$\mathrm{SNR}(d) = M \frac{GP_m}{N_0 W d^2} \tag{4.5}$$

is greater than or equal to 0 dB, where A_c is the area occupied by each cluster, λ is the carrier wavelength, P_m is the average transmit

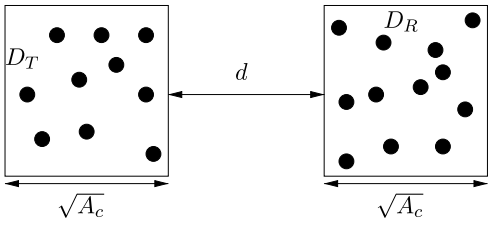

Fig. 4.1 Two square clusters of area A_c separated by distance d.

power per node during the MIMO transmissions, and d is the distance between the two clusters, assumed to be greater than or equal to $\sqrt{A_c}$ (Figure 4.1).

In Section 4.4, we state the above result in a precise manner and provide an intuitive argument for its proof. Meanwhile, let us explore what consequences it has on the throughput scaling of wireless networks.

4.3.2 Optimal Schemes

Capitalizing on the result of the previous section, we now complete the proof of Theorem 4.2, dividing the analysis into the two regimes $\sqrt{A}/\lambda \geq n$ and $\sqrt{n} \leq \sqrt{A}/\lambda < n$.

Full Hierarchical Cooperation ($\sqrt{A}/\lambda \geq n$). In this regime, no spatial limitation is felt at the network level and the upper bound in (4.3) allows for throughput scaling linear in n. Potentially hierarchical cooperation can achieve arbitrarily close to linear scaling. One needs to check, however, that the MIMO transmissions taking place at all levels of the scheme are fully efficient, i.e., have capacity scaling linearly in the number of nodes in the clusters. Consider therefore a MIMO transmission between two clusters of size M and area $A_c = AM/n$. In the hierarchical cooperation scheme, these two clusters are part of a larger cluster of size M' and area $A'_c = A_c M'/M$ in the next level of the hierarchy. The separation d between the two clusters is therefore upper bounded by

$$d \leq \sqrt{A'_c} = \sqrt{A_c M'/M}.$$

This in turn implies that the ratio $A_c/\lambda d$ is lower bounded by

$$\frac{A_c}{\lambda d} \geq \frac{A_c}{\lambda\sqrt{A_c M'/M}} = \frac{\sqrt{A_c M/M'}}{\lambda}.$$

As $\sqrt{A}/\lambda \geq n$ by assumption, we obtain that $\sqrt{A_c}/\lambda = (\sqrt{AM/n})/\lambda \geq \sqrt{nM}$, which implies that

$$\frac{A_c}{\lambda d} \geq \sqrt{nMM/M'} \geq M,$$

as n is greater than or equal to M'. Hence, condition (4.4) is satisfied and MIMO transmissions at all the levels of the hierarchical cooperation scheme operate with full degrees of freedom, just like in the case of i.i.d. phases. These transmissions also have sufficient power, satisfying the power condition in (4.5). Recall from Section 2.4 that during the MIMO transmissions, we transmit with elevated power $P_m = M'P/M$ per node. We still satisfy the average power constraint P per node, because each cluster transmits only a fraction M/M' of the total time because of the time-division between the MIMO transmissions. Therefore, the SNR for the MIMO transmissions is given by

$$\mathrm{SNR}(d) = M\frac{GP_m}{N_0 W A_c'} = M'\frac{GP}{N_0 W A_c'} = n\frac{GP}{N_0 W A} > 0\,\mathrm{dB}.$$

where the last inequality follows from the power condition $\mathrm{SNR}_l = nGP/(N_0 WA) > 0\,\mathrm{dB}$ for the network in Theorem 4.2. Therefore, MIMO transmissions at each level of the hierarchy have full degrees of freedom and sufficient power. Hierarchical cooperation achieves an aggregate throughput scaling arbitrarily close to linear in n in this case.

Hierarchical Cooperation under Spatial Limitation ($\sqrt{n} \leq \sqrt{A}/\lambda < n$). In this regime, Equation (4.3) shows that a linear throughput scaling is not achievable by any means. Nevertheless, the question remains whether one could outperform multi-hopping strategies, whose asymptotic performance $\Theta(\sqrt{n})$ is strictly suboptimal compared with the upper bound $O(\sqrt{A}/\lambda)$. A direct application of the hierarchical cooperation scheme fails to improve on multi-hop in this case; however, it turns out that a simple adaptation of the scheme to this spatially limited situation achieves the optimal scaling.

The idea is the following: organize the communication of the n source-destination pairs into n/N sessions, each involving N source-destination pairs, where $N = \sqrt{A}/\lambda$. It is possible to choose here the nodes in a way such that each group of N nodes statistically occupies the total area of the network. This way, no group of N nodes considered alone feels the spatial limitation, as for this diluted network $N = \sqrt{A}/\lambda$ and we are in the first case above. The sessions operate successively and the traffic in each session is handled using hierarchical cooperation where only the N chosen nodes are involved. The rest of nodes remain silent. As nodes are active only a fraction of N/n of the total time, when active they can transmit with elevated power $P_m = nP/N$ and still satisfy their individual power constraint P. Therefore, for the diluted network of N nodes in each session, the long-range SNR is

$$N\frac{GP_m}{N_0WA} = n\frac{GP}{N_0WA} > 0\,\text{dB}.$$

Therefore, the diluted network is neither power nor space-limited and hierarchical cooperation achieves aggregate throughput of order $N^{1-\varepsilon} = (\sqrt{A}/\lambda)^{1-\varepsilon}$ for any fixed $\varepsilon > 0$. With time-division across different groups of nodes, the same throughput is achievable in the whole network. This completes the proof of Theorem 4.2. □

4.4 Analysis of the Spatial Degrees of Freedom

We first reformulate precisely condition (4.4), as well as the claim made in Section 4.3.1, and then provide the key ideas behind the proof of this claim. The complete proof can be found in Ref. [27].

Lemma 4.3. Consider two square clusters of area A_c separated by distance d (see Figure 4.1), with each cluster containing M nodes distributed uniformly at random over A_c. Let $\sqrt{A_c} \ll d \ll A_c$, and let the nodes in the transmit cluster D_T perform independent signalling with power P each, such that the long-distance SNR between these two clusters defined as

$$\text{SNR}(d) = M\frac{GP}{N_0W\,d^2}$$

is greater than or equal to 0 dB. Then, there exists a constant $K > 0$ independent of M, A_c and d such that the capacity of the MIMO channel from the transmitting cluster D_T to the receiving cluster D_R is lower bounded by $C_{\mathrm{MIMO}} \geq KM$ with high probability as M grows large, as soon as

$$\frac{A_c/(\lambda d)}{\log(A_c/(\lambda d))} \geq KM. \tag{4.6}$$

Proof Idea. Under the above assumptions, the capacity of the MIMO channel can be expressed as

$$C_{\mathrm{MIMO}} = \log\det\left(I + \frac{\mathrm{SNR}(d)}{M} FF^\dagger\right),$$

where $\mathrm{SNR}(d)$ is assumed to be greater than or equal to 0 dB, and F is the rescaled $M \times M$ channel matrix whose entries are given by

$$F_{ik} = \frac{d}{\sqrt{G}} H_{ik} = \frac{d}{r_{ik}} \exp(2\pi j r_{ik}/\lambda).$$

The entries of F have therefore a magnitude of order 1, as $d \leq r_{ik} \leq d + o(d)$.

The first observation to be made is that the above capacity does not fluctuate much with respect to the random placement of the nodes. More precisely, it can be shown using standard concentration techniques that for any $t > 0$,

$$\mathbb{P}(|C_{\mathrm{MIMO}} - \mathbb{E}(C_{\mathrm{MIMO}})| > tM^{1/2+\varepsilon}) \leq \mathrm{e}^{-2t^2 M^{2\varepsilon}}.$$

What remains therefore to be shown is that under condition (4.6),

$$\mathbb{E}(C_{\mathrm{MIMO}}) = \Omega(M).$$

In order to obtain a lower bound on this expression, we use the same technique as in Section 2.4.2 and first show that there exists $K > 0$ such that

$$\mathbb{E}(C_{\mathrm{MIMO}}) \geq K \frac{M^4}{\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger))}. \tag{4.7}$$

The proof of this statement is relegated to Appendix A towards the end of the present section. Let us explore here the consequences of this inequality. As observed above, $|F_{ik}| \simeq 1$, so

$$\begin{aligned}\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)) &= \sum_{i,k,l,m=1}^{M} \mathbb{E}(F_{ik}F_{lk}^* F_{lm}F_{im}^*),\\ &= \sum_{i,k,m=1}^{M} \mathbb{E}(|F_{ik}|^2|F_{im}|^2) + \sum_{\substack{i,k,l=1\\ i\neq l}}^{M} \mathbb{E}(|F_{ik}|^2|F_{lk}|^2)\\ &\quad + \sum_{\substack{i,k,l,m=1\\ i\neq l,k\neq m}}^{M} \mathbb{E}(F_{ik}F_{lk}^* F_{lm}F_{im}^*),\\ &\simeq 2M^3 + \sum_{\substack{i,k,l,m=1\\ i\neq l,k\neq m}}^{M} \mathbb{E}(F_{ik}F_{lk}^* F_{lm}F_{im}^*).\end{aligned}$$

It follows that $\Theta(M^3) \leq \mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)) \leq \Theta(M^4)$, and correspondingly, $\Theta(1) \leq \mathbb{E}(C_{\mathrm{MIMO}}) \leq \Theta(M)$. The best scenario occurs when the random variables F_{ik} are centered and i.i.d., which is the situation encountered in Section 2; in this case, the four-sums term above vanishes; hence,

$$\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)) = O(M^3), \quad \mathbb{E}(C_{\mathrm{MIMO}}) = \Omega(M),$$

and the MIMO communication benefits from the full spatial degrees of freedom. In our case, $F_{ik} \simeq \exp(2\pi j r_{ik}/\lambda)$; hence,

$$\begin{aligned}&\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger))\\ &\quad \simeq 2M^3 + \sum_{\substack{i,k,l,m=1\\ i\neq l,k\neq m}}^{M} \mathbb{E}(\exp(2\pi j(r_{ik} - r_{lk} + r_{lm} - r_{im})/\lambda)).\end{aligned}$$

If it was the case that all points were aligned on the same horizontal line, then the phase shifts would cancel completely: $r_{ik} - r_{lk} + r_{lm} - r_{im} = 0$, and this would lead to

$$\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)) = O(M^4), \quad \mathbb{E}(C_{\mathrm{MIMO}}) = \Omega(1),$$

which would mean no spatial degrees of freedom for the MIMO transmission.[2]

In between these two extreme cases, where phases rotate independently, on the one hand, and are linearly dependent, on the other hand, lies the two-dimensional situation. We present a simplified argument below, showing that when condition (4.6) is met, the phase shifts generated by the random node placement are sufficiently important to ensure that the above four sums term remains small, providing therefore enough spatial degrees of freedom for the MIMO transmission.

The distance r_{ik} between two points $\mathbf{x}_i$ in D_T and $\mathbf{y}_k$ in D_R is given by

$$r_{ik} = \|\mathbf{x}_i - \mathbf{y}_k\| = \sqrt{(d + \sqrt{A_c}(x_{i1} + y_{k1}))^2 + A_c(x_{i2} - y_{k2})^2},$$

where the chosen coordinate system is illustrated in Figure 4.2; in particular, the horizontal and vertical coordinates x_{i1}, y_{k1}, x_{i2}, y_{k2} are rescaled so as to lie in the interval $[0,1]$ each. Using the assumption that $d \gg \sqrt{A_c}$, we obtain the following approximation[3]:

$$r_{ik} \simeq d + \sqrt{A_c}(x_{i1} + y_{k1}) + \frac{A_c}{2d}(x_{i2} - y_{k2})^2.$$

Based on this approximation, a short algebraic computation shows that

$$r_{ik} - r_{lk} + r_{lm} - r_{im} \simeq -\frac{A_c}{d}(x_{l2} - x_{i2})(y_{m2} - y_{k2}).$$

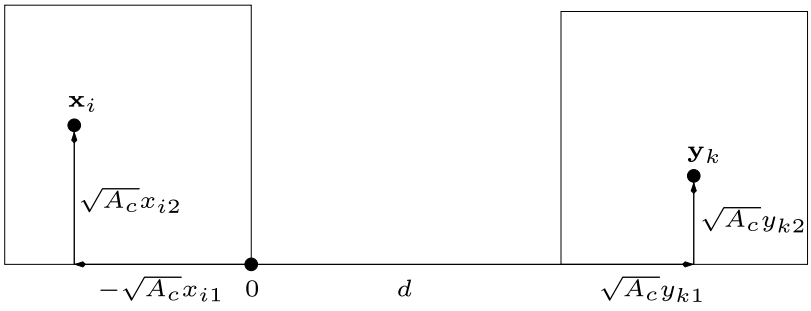

Fig. 4.2 Coordinate system.

[2] Technically speaking, as Equation (4.7) only provides a *lower bound* on $\mathbb{E}(C_{\mathrm{MIMO}})$, a further check would be required in order to conclude. It turns out that the MIMO capacity is indeed $\Theta(1)$ in this case.

[3] Let us point out here that the validity of this approximation is subject to caution. Nevertheless, it provides an intuitive argument justifying the existence of condition (4.6).

We see here that due to the differences in the vertical coordinates, the sum of these four terms is not equal to zero, contrary to the one-dimensional case mentioned above. The computation of the expectation gives

$$\begin{aligned}\mathbb{E}(\exp(2\pi j(r_{ik} - r_{lk} + r_{lm} - r_{im})/\lambda)) \\ \simeq \int_0^1 \mathrm{d}x_{i2} \int_0^1 \mathrm{d}x_{l2} \int_0^1 \mathrm{d}y_{k2} \int_0^1 \mathrm{d}y_{m2} \\ \times \exp\left(-2\pi j \frac{A_c}{\lambda d}\,(x_{l2} - x_{i2})\,(y_{m2} - y_{k2})\right). \end{aligned} \tag{4.8}$$

This multiple integral is shown below to be upper bounded by

$$K \frac{\lambda d}{A_c} \log\left(\frac{A_c}{\lambda d}\right),$$

which shows in turn that if condition (4.6) is met, then

$$\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)) \leq 2M^3 + M^4 K \frac{\lambda d}{A_c} \log\left(\frac{A_c}{\lambda d}\right) = O(M^3).$$

Combining this with (4.7) shows that $\mathbb{E}(C_{\mathrm{MIMO}}) = \Omega(M)$; the MIMO transmission is therefore fully efficient in this case; this completes the proof of Lemma 4.3. □

Proof of the Lower Bound (4.7). The proof follows the lines of Section 2.4.2: We have

$$\begin{aligned}\mathbb{E}(C_{\mathrm{MIMO}}) &= \mathbb{E}\left(\log\det\left(I + \frac{\mathrm{SNR}(d)}{M} FF^\dagger\right)\right), \\ &= M\mathbb{E}(\log(1 + \mathrm{SNR}(d)\lambda)) \geq M \log(1 + \mathrm{SNR}(d)t)\,\mathbb{P}(\lambda > t),\end{aligned}$$

where $t > 0$ and λ is an eigenvalue of $(1/M)FF^\dagger$, picked uniformly at random. By Paley–Zygmund's inequality (see Appendix B), we obtain that for $0 < t < \mathbb{E}(\lambda)$,

$$\mathbb{E}(C_{\mathrm{MIMO}}) \geq M \log(1 + \mathrm{SNR}(d)t) \frac{(\mathbb{E}(\lambda) - t)^2}{\mathbb{E}(\lambda^2)}.$$

As observed above, the entries of F are of order 1; hence,

$$\mathbb{E}(\lambda) = \frac{1}{M}\mathbb{E}\left(\mathrm{Tr}\left(\frac{1}{M}FF^\dagger\right)\right) = \frac{1}{M^2}\sum_{i,k=1}^{M} \mathbb{E}(|F_{ik}|^2) \simeq 1.$$

In addition,

$$\mathbb{E}(\lambda^2) = \frac{1}{M}\mathbb{E}\left(\mathrm{Tr}\left(\left(\frac{1}{M}FF^\dagger\right)^2\right)\right) = \frac{1}{M^3}\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger)),$$

hence, choosing finally $t = 1/2$ in the above estimate gives

$$\mathbb{E}(C_{\mathrm{MIMO}}) \geq \frac{\log(1+\mathrm{SNR}(d)/2)}{4}\frac{M^4}{\mathbb{E}(\mathrm{Tr}(FF^\dagger FF^\dagger))},$$

which proves the claim. □

Computation of the Multiple Integral (4.8). The computation of the innermost integral gives

$$\begin{aligned}\int_0^1 \mathrm{d}y_{m2} \exp&\left(-2\pi j\frac{A_c}{\lambda d}(x_{l2}-x_{i2})(y_{m2}-y_{k2})\right)\\ &= -\frac{\lambda d}{2\pi j A_c(x_{l2}-x_{i2})}\\ &\quad\times\exp\left(-2\pi j\frac{A_c}{\lambda d}\left(x_{l2}-x_{i2}\right)\left(y_{m2}-y_{k2}\right)\right)\Big|_{y_{m2}=0}^{y_{m2}=1},\end{aligned}$$

implying that

$$\left|\int_0^1 \mathrm{d}y_{m2}\exp\left(-2\pi j\frac{A_c}{\lambda d}(x_{l2}-x_{i2})(y_{m2}-y_{k2})\right)\right| \leq K\frac{\lambda d}{A_c}\frac{1}{|x_{l2}-x_{i2}|},$$

for a constant $K > 0$ independent of A_c and d. Dividing the integral over the x's into two domains $|x_{l2}-x_{i2}| \leq \varepsilon$ and $|x_{l2}-x_{i2}| > \varepsilon$, we therefore obtain

$$\begin{aligned}&|\mathbb{E}(\exp(2\pi j(r_{ik}-r_{lk}+r_{lm}-r_{im})/\lambda))|\\ &\quad\leq 2\varepsilon + 2\int_0^{1-\varepsilon}\mathrm{d}x_{i2}\int_{x_{i2}+\varepsilon}^1 \mathrm{d}x_{l2}K\frac{\lambda d}{A_c}\frac{1}{|x_{l2}-x_{i2}|}\\ &\quad\leq 2\varepsilon + K\frac{\lambda d}{A_c}\log(1/\varepsilon).\end{aligned}$$

As this upper bound is valid for any $0 < \varepsilon < 1$, choosing $\varepsilon = \lambda d/A_c$ leads to the desired result:

$$|\mathbb{E}(\exp(2\pi j(r_{ik}-r_{lk}+r_{lm}-r_{im})/\lambda))| \leq K\frac{\lambda d}{A_c}\log\left(\frac{A_c}{\lambda d}\right).$$ □

A

Regularity Properties of Random Networks

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

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

if

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

The regularity properties given below arise from the assumption that nodes are distributed uniformly at random over the network area and source-destination pairings are also formed randomly without any consideration of node locations.

Lemma A.1. The random network of n nodes with area A and random source-destination pairings satisfies the following properties:

(a) Let us partition the network area A into cells of area A_c, where A_c can be a function of n and A. Then for any

$0 < \delta < 1$, the number of nodes inside each cell is in the interval $((1-\delta)(A_c/A)n, (1+\delta)(A_c/A)n)$ with probability larger than $1 - (2A/A_c)\mathrm{e}^{-\Lambda(\delta)(A_c/A)n}$, where $\Lambda(\delta)$ is independent of n, A and A_c, and satisfies $\Lambda(\delta) > 0$ when $\delta > 0$.

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

Remark A.2. Even though the condition is not explicitly mentioned above, part (a) of the above lemma is interesting only when

$$\frac{A_c n}{A} \geq n^{\gamma} \quad \text{for some } \gamma > 0,$$

which does not include the case where, for example, A scales like n and A_c is constant.

Proof of Lemma A.1.

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

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{n} B_i \geq (1+\delta)\frac{A_c}{A}n\right) \\
&\quad = \mathbb{P}(\mathrm{e}^{s\sum_{i=1}^{n} B_i} \geq \mathrm{e}^{s(1+\delta)(A_c/A)n}) \qquad \text{(A.1)} \\
&\quad \leq (\mathbb{E}[\mathrm{e}^{sB_1}])^n \mathrm{e}^{-s(1+\delta)(A_c/A)n} \\
&\quad = \left(\mathrm{e}^{s}\frac{A_c}{A} + \left(1 - \frac{A_c}{A}\right)\right)^n \mathrm{e}^{-s(1+\delta)(A_c/A)n} \\
&\quad \leq \mathrm{e}^{(A_c/A)n(e^s-1)}\, e^{-s(1+\delta)(A_c/A)n} \\
&\quad = \mathrm{e}^{-(A_c/A)n\Lambda_+(\delta)} \qquad \text{(A.2)}
\end{aligned}$$

by choosing $s = \ln(1+\delta)$, where $\Lambda_+(\delta) = (1+\delta)\ln(1+\delta) - \delta$. Note that $\Lambda_+(\delta) > 0$ when $\delta > 0$. The probability of

having a cell with more than $(1+\delta)(A_c/A)n$ nodes is upper bounded by the union bound as

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

The proof for the lower bound follows similarly and yields

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

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

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

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

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

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

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

B

The Paley–Zygmund Inequality

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

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

Proof of Lemma B.1. By the Cauchy–Schwarz inequality, we have for any $t \geq 0$

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

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

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

Therefore,

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

References

[1] S. Ahmad, A. Jovicic, and P. Viswanath, "Outer bounds to the capacity region of wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2770–2776, June 2006.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*. Cambridge University Press, 2009.

[3] V. Cadambe and S. Jafar, "Interference alignment and the degrees of freedom for the K-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, August 2008.

[4] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, September 1979.

[5] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *AT&T Bell Labs Tech. Journal*, vol. 1, no. 2, pp. 41–59, 1996.

[6] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 1009–1018, March 2007.

[7] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3413–3424, August 2009.

[8] J. Ghaderi, L.-L. Xie, and X. Shen, "Hierarchical cooperation in ad hoc networks: Optimal clustering and achievable throughput," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3425–3436, August 2009.

[9] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 388–404, March 2000.

[10] A. Jovicic, P. Viswanath, and S. R. Kulkarni, "Upper bounds to transport capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2555–2565, November 2004.

[11] S. Katti, S. Gollakota, and D. Katabi, "Embracing wireless interference: Analog network coding," in *Proceedings of the ACM SIGCOMM*, Kyoto, Japan, August, 2007.

[12] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, September 2005.

[13] G. Kramer, I. Maric, and R. Yates, "Cooperative Communications," *Foundations and Trends in Networking*, vol. 1, no. 3–4, pp. 271–425, 2006.

[14] S.-H. Lee and S.-Y. Chung, "Effect of channel correlation on the capacity scaling in wireless networks," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Toronto, Canada, July, 2008.

[15] S.-H. Lee and S.-Y. Chung, "On the capacity scaling of wireless ad hoc networks: Effect of finite wavelength," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Austin, TX, July, 2010.

[16] O. Lévêque and E. Telatar, "Information theoretic upper bounds on the capacity of large, extended ad-hoc wireless networks," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 858–865, March 2005.

[17] J. Liu, D. Goeckel, and D. Towsley, "Bounds on the throughput gain of network coding in unicast and multicast wireless networks," *IEEE Transactions on Selected Areas in Communications*, vol. 27, no. 5, pp. 582–592, May 2009.

[18] B. Nazer, M. Gastpar, S. Jafar, and S. Vishwanath, "Ergodic interference alignment," in *Proceedings of the IEEE International Symposium on Information Theory*, Seoul, South Korea, July, 2009.

[19] U. Niesen, "Interference alignment in dense wireless networks," *IEEE Transactions on Information Theory*, vol. 57, no. 5, pp. 2889–2901, May 2011.

[20] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 3959–3982, September 2009.

[21] U. Niesen, P. Gupta, and D. Shah, "The balanced unicast and multicast capacity regions of large wireless networks," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2249–2271, May 2010.

[22] A. Özgür, "Fundamental limits and optimal operation in large wireless networks," PhD Thesis Nr 4483, EPFL, 2009.

[23] A. Özgür, R. Johari, O. Lévêque, and D. Tse, "Information theoretic operating regimes of large wireless networks," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 427–437, January 2010.

[24] A. Özgür and O. Lévêque, "Throughput-delay tradeoff for hierarchical cooperation in ad hoc wireless networks," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1369–1377, March 2010.

[25] A. Özgür, O. Lévêque, and E. Preissmann, "Scaling laws for one and two-dimensional random wireless networks in the low attenuation regime," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3573–3585, October 2007.

[26] A. Özgür, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad-hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3549–3572, October 2007.

[27] A. Özgür, O. Lévêque, and D. Tse, "Linear capacity scaling in wireless networks: Beyond physical limits?," in *Proceedings of the IEEE Information Theory and Applications Workshop*, San Diego, CA, February, 2010.

[28] A. Özgür and D. Tse, "Achieving linear scaling with interference alignment," in *Proceedings of the IEEE International Symposium on Information Theory*, Seoul, South Korea, 2009.

[29] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–596, November 1999.

[30] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[31] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communications: Scaling laws and optimal operation," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 748–767, May 2004.

[32] L.-L. Xie and P. R. Kumar, "On the path-loss attenuation regime for positive cost and linear scaling of transport capacity in wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2313–2328, June 2006.

[33] F. Xue, L.-L. Xie, and P. R. Kumar, "The transport capacity of wireless networks over fading channels," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 834–847, March 2005.

[34] S. Zhang, S. Liew, and P. Lam, "Physical layer network coding," in *Proceedings of the ACM MobiCom*, Los Angeles, CA, September, 2006.