UNIVERSITY OF MINNESOTA

This is to certify that I have examined this copy of a doctoral Dissertation by

Alejandro Ribeiro

and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

Name of Faculty Advisor(s)

Signature of Faculty Advisor(s)

Date

GRADUATE SCHOOL

Wireless Cooperative Communications and Networking

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Alejandro Ribeiro

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Georgios B. Giannakis, Advisor

December 2006

"We knew so little then. I know even less now."

Acknowledgments

Of the countless hours spent writing this thesis the few minutes I will devote to write the next paragraphs are among the most enjoyable. This journey I call my life is no more than the story of the people I have met along its path, and it is not everyday that I take some time to remember how special these individuals are.

This thesis is the result of fruitful collaboration with my Ph. D. thesis advisor Prof. Georgios Giannakis. Though not always politely, we have mutually challenged each other, my most cherished results stemming from these sometimes heated discussions. I thank him for the many times he told me I was wrong, and for allowing me to reciprocate even when he was not. More important than that, I thank him for offering his friendship.

Due thanks go to Dr. Ananthram Swami and Profs. Mos Kaveh, Zhi-Quan (Tom) Luo, Stergios Roumeliotis, Nikos Sidiropoulos, and Zhi-Li Zhang for agreeing to serve on my committee. I am doubly grateful to Dr. Swami and Prof. Sidiropoulos for traveling to attend my Ph. D. thesis defense.

The work in this thesis would not have been possible without the help of the people with whom I have collaborated in the last three years that deserve not only my gratitude but due credit for their significant contributions to the work reported here. A plead to accept my gratitude is thus extended to Prof. Xiaodong Cai, Dr. Alfonso Cano-Pleite, Prof. Zhi-Quan (Tom) Luo, Prof. Stergios Roumeliotis, Ioannis Schizas, Ali Faisal Sha, Prof. Nikos Sidiropoulos, Dr. Renqiu Wang, Tairan Wang, Prof. Xin Wang, Dr. Jinjun Xiao, and Yingqun Yu. The material here has also benefited from discussions with current and former lab-mates at the University of Minnesota: Juan Andres Bazerque, Shahrokh Farahmand, Antonio García Marques, Vikrham Gowreesunker, Gonzalo Mateos, Eric Msechu, Dr. Qingwen Liu, Dr. Xiliang Luo, Dr. Pengfei Xia, Prof. Liuqing Yang, and Dr. Wanlun Zhao. To those of you that concede me the honor of your friendship I must say that even if you could not make my days longer, you have made them better.

When I was looking for a way to finance my Ph. D. degree during the Uruguayan winter of 2002 I counted on the help of a number of people that believed in me without any good reason. They are Dr. Lucía Colombino and Dr. Mercedes Jiménez de Aréchaga from the Fulbright Commission in Uruguay, Dr. Gregory Randall from the Universidad de la República Oriental del Uruguay, and Dr. Guillermo Sapiro from the University of Minnesota. Upon arriving to Minnesota I counted with the invaluable help of my good friends Dr. Alberto Bartesaghi and Dr. Facundo Memoli. None of this would have been

possible without your help.

Por último también tengo que agradecer a todos ustedes que no tienen nada que ver con esta tesis pero todo que ver con mi vida. Papá, Mamá y Nelson, gracias por ser mi familia. La vida es tan larga y el mundo tan pequeño, nos vemos pronto. Y a mis amigos que están tan lejos, Rodrigo Arizaga, Rafael Bernardi, Rafael Grompone, Federico Lecumberry, Álvaro Pardo, Adriana Piazza y Aiala Rosá, gracias por estar tan cerca.

Y para vos Gabriela pensaba dedicarte esta tesis pero no me pareció suficiente. Quizás el experimento que estamos haciendo juntos sí sea suficiente. Para Miranda y Guille.

Alejandro Ribeiro
Minneapolis, November 12 2006.

Abstract

Pathloss and fading are unique features of wireless propagation, respectively referring to the rapid decay in the received signal envelope with distance and to the random fades present in the received signal power. Multi-hopping and diversity are the corresponding countermeasures entailing the division of a longer link into shorter links and the provision of diversified information bearing signal replicas at the destination. This thesis builds on the fact that by letting users collaborate in relaying packets for each other they can obtain independent propagation paths to reach their intended destinations through a series of shorter hops; thus mitigating both pathloss and fading. In a nutshell, collaboration offers both diversity and multi-hopping benefits at the same time. This thesis consists of two interrelated thrusts which explore the role of user collaboration in multiple access networks as a diversity enabler and the role of multihop routing in counteracting the rapid decrease in average received power. Our results suggest that joint exploitation of multipath and multi-hop links in the context of collaborative networking offers substantial improvement in terms of capacity, coverage, power consumption and error performance. Even though different in the principles they exploit, both thrusts commonly rely on what we purport as a paradigm shift in wireless networks: from competition towards collaboration.

We show that user cooperation in random access networks (RA) yields a significant increase in throughput. Specifically, we prove that for networks with a large number of users, the throughput of a cooperative wireless RA network operating over Rayleigh fading links approaches the throughput of an RA network operating over additive white Gaussian noise links. The message borne out of this result is that user cooperation offers a viable choice for migrating diversity benefits to the wireless RA regime, thus bridging the gap to wireline RA networks, without incurring a bandwidth or energy penalty.

In the context of multi-hop routing, existing graph-theoretic approaches rely on so-called disk models. Albeit valuable for wired networks, these models do not capture adequately the random nature of wireless links. To this end, we introduce a novel framework for stochastic routing in wireless multihop networks, whereby each node selects a neighbor to forward a packet with a certain probability. A plethora of valuable criteria emerge from this framework based on which these routing probabilities are obtained efficiently as solutions of typically convex optimization problems. We further develop distributed self-organizing stochastic routing via primal dual decomposition solvers, and study the associated convergence properties.

Contents

List of Figures

Chapter 1

From competition towards collaboration

The importance of wireless communication networks in everyday life is difficult to overstress. Besides the now ubiquitous cell-phones, wireless-enabled computers and the envisioned proliferation of wireless sensors for monitoring and surveillance render wireless networks vital to the growth of as diverse sectors as environmental, financial, healthcare, and manufacturing. However, conventional wireless networks are operating at or close to their limits, generating a recent research spur in disruptive technologies like cognitive radio, collaborative networking and wireless sensor networks, to name a few. The focus of this dissertation is in what we purport as a paradigm shift in wireless networking: from competition towards collaboration.

The existing wireless networking paradigm consists of groups of user terminals communicating with any out of a group of access points (APs). The APs may be cellular base stations deployed in a certain city where the user terminals represent wireless phones; or they may represent 802.11 APs in a building with the user terminals denoting wireless enabled computers. In one form or the other, existing protocols entail contention of user terminals to access limited resources offered by the APs. In the coverage area of an 802.11 local area network, terminals (randomly) contend to reach the AP; in the GSM cellular standard they compete for time slots; whereas in the IS-95 CDMA standard the

constrained resource comprises a number of spreading codes available and a minimum acceptable signal-to-interference-plus-noise ratio. Despite arguably major differences, a star topology is common to all these settings and collaboration among terminals goes no farther than controlling mutual interference.

Replacing competition with collaboration is at the midst of many research areas in wireless networking. Indeed, collaboration between terminals is the per-se enabler of mesh and ad-hoc networks whereby terminals collaborate in relaying packets for each other. In this way, terminals convey their information bearing signals via a route of lesser power consumption, while at the same time improving resilience against channel fades.

In a nutshell this thesis concerns to the theory and implementation of wireless collaborative networks. The reader is surely aware of somewhat independent bodies of knowledge pertaining to wireless communications and (wired) networking. She or he thus may rightfully wonder why these results are not directly applicable. The short answer is that they are, but yield largely suboptimal performance in general. Indeed, the properties of a network are not direct extensions of the properties of individual links, thus implying that results pertaining to wireless point-to-point channels do not apply verbatim to wireless networks. On the other hand, wireless networks are subtly yet fundamentally different from their wired counterparts. The reasons for this are many but on a basic level they all reduce to the fact that in lieu of a tangible connection, a link in wireless network is somewhat of an arbitrary definition. In this chapter we present an overview of known results in networking and wireless communications as required for the material covered in the rest of the thesis.

1.1 Wireless channels

At first sight it may not be clear why there should be such a thing as a wireless communication channel requiring an approach different than the one used for ordinary (wired) channels. Surely enough, the physical phenomena explaining a radio channel and, say, signal propagation over copper wires are different. After all, even if Albert Einstein never said it, it is indeed true that “The ordinary telegraph is like a very long cat. You pull the tail in New York, and it meows in Los Angeles. The wireless is the same, only without the cat.”.

But equally different are the physical phenomena involved in a copper wire and an optical fiber, not preventing communication system designers to use similar models and techniques.

In fact, copper wires and optical fibers share properties that bring them close from a communications perspective. A wireless channel, on the other hand, exhibits three features that render it fundamentally different: pathloss, fading, and broadcast propagation[1].

To be precise, consider a source S transmitting an L symbol packet $\mathbf{s}_S := [s_S(0), \dots s_S(L-1)]$. The symbols are linearly modulated by a unit energy pulse $p(t)$ of duration T_s so that the signal transmitted by S is

$$x_S(t) = \sum_{l=0}^{L-1} s_S(l) p(t - lT_s). \tag{1.1}$$

This $x_S(t)$ waveform propagates through the channel and is eventually received at destination D. Denoting by $h(t)$ the transference of the channel, we have that the received signal $z_D(t)$ is given by

$$z_D(t) = h_{DS}(t) x_S(t) + n(t) \tag{1.2}$$

where $n(t)$ denotes additive white Gaussian noise (AWGN) with double sided spectral density $N_0/2$.

For simplicity, and consistent with the treatment throughout the thesis, we assume that $h_{DS}(t)$ is known at the receiver side D – in practice it is estimated. Consequently, the optimal receiver front-end is a filter matched to the equivalent pulse $h(t)p(t-lT_s)$. In other words the optimal receiver constructs the discrete time signal

$$y_D(l) = \int_{lT_s}^{(l+1)T_s} z_D(t) \frac{[h(t)p(t - lT_s)]^*}{\|h(t)p(t - lT_s)\|}\, dt \tag{1.3}$$

and utilizes the sequence $\mathbf{y}_D := [y_D(0), \dots y_D(L-1)]$ to compute the optimal estimate $\hat{\mathbf{s}}_S$ of $\mathbf{s}_S$. Upon defining the channel $h_{DS}(l) = \int_{lT_s}^{(l+1)T_s} \|h(t)p(t - lT_s)\|^2 dt$ we can write the equivalent discrete-time baseband system as

$$y_D(l) = h_{DS}(l) s_S(l) + n(l) \tag{1.4}$$

where direct computations show that the noise power is $\mathrm{E}[n^*(l)n(l)] = N_0$.

[1]There is a fourth feature called shadowing that we do not consider in this thesis.

With the model in (1.4) characterizing any communication systems in which inter-symbol interference can be neglected, the difference between wireless channels and wired channels is in the model of the channel transference $h(l)$. Whereas in a wired channel it is possible to assume that $h_{DS}(l)$ is a known constant in a wireless channels the transference $h_{DS}(l)$ exhibits random variations. Depending on how fast these variations are with respect to the packet duration LT_s we encounter different challenges, motivating the definition of so called fast, block, and slow varying fading channels that we discuss in the next section.

1.1.1 Fast, block, and slow varying fading

As the electromagnetic wave emitted by S propagates it is scattered, reflected, and diffracted in the many buildings, vehicles, persons, and geographical landmarks that define the physical environment that contains S and D. Ultimately, the interference generated by the various propagation paths creates a stationary wave with a certain interference pattern. While it is intractable to compute this interference pattern, we can infer two things about it: i) there is a large number of propagation paths contributing to the signal amplitude and phase at a certain position; and ii) the maxima and minima of this interference pattern are spaced by c/ν, where c denotes the wave speed and ν the transmitter frequency.

That there exists a large number of paths contributing to the channel transference justifies the customary assumption that the random variations of $h(t)$ have a complex Gaussian distribution by appealing to a central limit theorem argument. In turn, this implies that the fading coefficient $h_{DS}(l)$ in the discrete time model in (1.4) can be approximately modeled as the absolute value of this complex Gaussian random variable. With $\mathrm{E}(\cdot)$ denoting mathematical expectation, the distribution of the squared envelope is exponential; see e.g. [69, Ch. 2]

$$f_{h_{DS}^2}(h_{DS}^2) = \frac{1}{\mathrm{E}(h_{DS}^2)} e^{-h_{DS}^2/\mathrm{E}(h_{DS}^2)}. \tag{1.5}$$

The randomness of the channel dictates that it is important to distinguish between instantaneous and average performance metrics. Consider, e.g., the signal-to-noise-ratio (SNR) and let $P_S := \mathrm{E}[s_S^2(l)]$ denote the average power transmitted by the source. Given a realization of the fading coefficient h_{DS}, the power received at D from S is $h_{DS}^2(l)P_S$ and the

Table 1.1: Doppler frequencies for usual frequencies and user velocities.

	5 km/h (pedestrian)	40 km/hr (street vehicle)	100 km/hr (highway vehicle)
900 MHz (GSM)	3.7 Hz	29.6 Hz	74.0 Hz
1.9 GHz (PCS)	8.8 Hz	70.4 Hz	175.9 Hz
2.1 GHz (unlicensed)	9.7 Hz	77.8 Hz	194.4 Hz

corresponding *instantaneous* SNR is given by

$$\gamma_{DS}(l) = \frac{h_{DS}^2(l) P_S}{N_0}. \tag{1.6}$$

An alternative figure of merit is the *average* SNR which we define as the expected value of $\gamma_{DS}(l)$ in (1.6), namely

$$\bar{\gamma}_{DS} := \mathrm{E}[\gamma_{DS}(l)] = \frac{\mathrm{E}[h_{DS}^2(l)] P_S}{N_0}. \tag{1.7}$$

Which one of the two performance metrics in (1.6) and (1.7) is relevant to a particular problem is largely determined by how fast the variations in $h_{DS}(l)$ are with respect to a block duration, taking us back to the interference pattern.

As mentioned before, the interference pattern is characterized by the wave length c/ν in the sense that for points separated by this distance we are in a different lobe of the interference pattern. Thus, if we consider a terminal moving with velocity v we can roughly assess the rate of channel variations as $\nu v/c$, implying that in a unit of time, we can expect $\nu v/c$ realizations of the channel h_{DS}. For usual frequency ranges and user velocities, this so called Doppler frequencies range from a few Hertz to a few hundred Hertz as we show in Table 1.1. We consider there the frequency bands used by the most popular wireless standards namely the global system for mobile communications (GSM) operating at around 900 MHz, the personal communications systems (PCS) band at 1.9 GHz and the unlicensed band at 2.1 GHz used by terminals in wireless local area networks (802.11x).

The relation between the Doppler frequency $\nu v/c$ and the packet duration given by LT_s determines how many fading instantiations a packet – or group of packets – experiences. This determines a classification of fading channels as follows:

Fast fading channels. This is the case when $LT_s\nu v/c \gg 1$ implying that every packet experiences a large number of fading states. Interestingly, error control codes can be used to average fading states through the packet, and fast fading channels behave approximately like AWGN channels; see e.g., [69, ch. 14]. In this case the relevant SNR metric is the average SNR $\bar{\gamma}_{DS}$ in (1.7).

Block fading channels. When $LT_s\nu v/c \approx 1$, the fading coefficient remains unchanged during the duration of a packet. In this case the instantaneous SNR does not depend on l and we write for convenience

$$\gamma_{DS}(l) = \gamma(h_{DS}^2) = \frac{h_{DS}^2 P_S}{N_0}. \tag{1.8}$$

For block fading channels, the instantaneous SNR determines the error probability of a given packet. The average error probability for a sequence of packets, as we will see, can be written in terms of the average SNR in (1.7).

Slow fading channels. When $LT_s\nu v/c \ll 1$, the fading coefficient is not only invariant over the duration of a packet but presumably for the whole length of the communication. In this case the instantaneous SNR in (1.8) is the metric of interest.

1.1.2 Error probability

The SNR is a relevant performance metric only to the extent that it determines the probability of correct detection, or conversely, the error probability. In an AWGN channel the symbol error probability (SEP) depends on the type of modulation used but in general it can be written (or at least bounded) in terms of the Gaussian tail function $Q(x) := (1/\sqrt{2\pi}) \int_x^\infty e^{-u^2/2}\, du$. In fact, for a large class of modulation alternatives the SEP is given by

$$q_e^G(\bar{\gamma}_{DS}) = Q\left[\sqrt{k\bar{\gamma}_{DS}}\right] \tag{1.9}$$

with a properly selected k. Error probability for, e.g., binary phase shift keying (BPSK) is obtained from (1.12) with $k = 2$.

To obtain the probability of error in decoding the block $\mathbf{s}_S$, we have to consider the error correcting code used. For illustration purposes consider a BCH block code capable of

correcting up to $\epsilon_{\max}$ errors [69, p.437]. These codes are such that a packet is incorrectly decoded only when more than $\epsilon_{\max}$ bits have been incorrectly decoded and consequently the packet error probability (PEP) for a Gaussian channel is given by

$$P_e^G(\bar{\gamma}_{DS}) = 1 - \sum_{\epsilon=0}^{\epsilon_{\max}} \binom{L}{\epsilon} q^{\epsilon}(\bar{\gamma}_{DS})[1 - q(\bar{\gamma}_{DS})]^{L-\epsilon}. \tag{1.10}$$

For a given realization of h_{DS} a fading channel is not different from an AWGN channel and we can thus use (1.10) to understand PEP in slow, fast, and block fading channels.

In a slow fading channel we are interested in a single fading realization. The PEP of interest is thus identical to (1.10) using the instantaneous $\gamma(h_{DS}^2)$ instead of $\bar{\gamma}_{DS}$; i.e.,

$$P_e^S[\gamma(h_{DS}^2)] = 1 - \sum_{\epsilon=0}^{\epsilon_{\max}} \binom{L}{\epsilon} q^{\epsilon}[\gamma(h_{DS}^2)]\big[1 - q[\gamma(h_{DS}^2)]\big]^{L-\epsilon}. \tag{1.11}$$

For a fast fading channel the bit error probability changes from bit to bit. In general we can write for the l-th bit

$$q_e^F[\gamma_{DS}(l)] = Q\left[\sqrt{k\gamma_{DS}(l)}\right]. \tag{1.12}$$

The packet error probability can then be obtained by considering all possible combinations of bit error sequences with more than 5 errors. While beyond the scope of this thesis, it can be seen that the PEP for fast fading channels $P_e^F[\bar{\gamma}_{DS}]$ can be approximated by the corresponding PEP for Gaussian channels $P_e^F[\bar{\gamma}_{DS}] \approx P_e^G[\bar{\gamma}_{DS}]$; see e.g., [69, Ch. 14].

For block fading channels, the probability of a given packet being incorrectly decoded coincides with that of a slow fading channel in (1.11), i.e., $P_e^B[\gamma(h_{DS}^2)] = P_e^S[\gamma(h_{DS}^2)]$. However, for a block fading channel the average error probability is also of interest and can be obtained by averaging $P_e^B[\gamma(h_{DS}^2)]$ over all possible fading realizations to obtain

$$\bar{P}_e^B(\bar{\gamma}_{DS}) = \int_0^{\infty} P_e^B[\gamma(h_{DS}^2)] f_{h_{DS}^2}(h_{DS}^2)\ . \tag{1.13}$$

Assuming ergodicity of the channel process, the average PEP $\bar{P}_e^B(\bar{\gamma}_{DS})$ in (1.13) can be interpreted as the probability of missing a packet when the communication $S \to D$ is observed during long periods of time. This has to be contrasted with $P_e^B[\gamma(h_{DS}^2)]$ that represents the error probability for a given packet. Alternatively, $\bar{P}_e^B(\bar{\gamma}_{DS})$ can be thought of as the error probability that S, not knowing h_{DS}, expects to experience when transmitting to D.

1.1.3 Pathloss and multi-hopping

Pathloss refers to the expected value $\mathrm{E}[h_{DS}^2(l)]$ of the channel transference determining the relation between the power P_S transmitted by a source terminal S and the power P_{DS} received at the intended destination D. Letting $\mathbf{S}$ and $\mathbf{D}$ denote the positions of S and D respectively, we have that when S transmits with power P_S, the power P_{DS} received at D from S is given by

$$P_{DS} = P_S L(\mathbf{D} - \mathbf{S}) \tag{1.14}$$

where $L(\mathbf{D} - \mathbf{S})$ is a distance-dependent pathloss coefficient. While different models are available for $L(\mathbf{D} - \mathbf{S})$, all of them predict a exponential decay of received power with distance of the form $L(\mathbf{D}-\mathbf{S}) = \xi\|\mathbf{D}-\mathbf{S}\|^{-\alpha}$. Typical values of α vary between 3 and 4 and the constant ξ depends on properties of the physical environment, e.g., the type of human development – industrial, rural, urban, etc. – and properties of the transmitter / receiver pair, e.g., the radiation pattern of transmitter and receiver antennas.

The effect of exponential pathloss in what pertains to this thesis is that reducing the distance between S and D has a clear potential to reduce power consumption at the source. To be precise assume that for guaranteeing a target error probability performance we require a given received power P_{DS}, and compare the power P_{S_0} that the source needs to transmit when its position is $\mathbf{S}_0$, with the power P_{S_1} when the position is $\mathbf{S}_1$. The power ratio is [cf. (1.14) with $L(\mathbf{D} - \mathbf{S}) = \xi\|\mathbf{D} - \mathbf{S}\|^{-\alpha}$]

$$\frac{P_{S_1}}{P_{S_0}} = \left(\frac{\|\mathbf{D} - \mathbf{S}_1\|}{\|\mathbf{D} - \mathbf{S}_0\|}\right)^{\alpha}. \tag{1.15}$$

If the distance between S and D is doubled, i.e., $\|\mathbf{D} - \mathbf{S}_1\| = 2\|\mathbf{D} - \mathbf{S}_0\|$, in order to maintain quality of service S has to increase its transmit power by a factor 2^{α}, and with a rather conservative $\alpha = 3.4$ this entails a tenfold increase in P_S. When the distance from S to D increases by a factor of 10, P_S increases three orders of magnitude by a factor $10^{3.4} = 2.5 \times 10^3$.

This provides a motivation to exploit user collaboration to reduce the average distance between communicating pairs of nodes. The idea is that instead of attempting direct transmission between S and D consuming power P_{S_0}, we divide the link in successive commu-

nications between S and a relay terminal R with power P_{S_1} followed by communication between R and D with power P_R. For a fair comparison we require the powers P_{RS_1} received at R from S and P_{DR} received at D from R in the relay assisted communication to coincide with the power P_{DS_0} received at D from S in the original direct link. This setup allows for a power reduction factor of

$$\frac{P_{S_1}+P_R}{P_{S_0}} = \left(\frac{\|\mathbf{R}-\mathbf{S}_1\|}{\|\mathbf{D}-\mathbf{S}_1\|}\right)^{\alpha} + \left(\frac{\|\mathbf{D}-\mathbf{R}\|}{\|\mathbf{D}-\mathbf{S}_1\|}\right)^{\alpha}. \tag{1.16}$$

Placing, e.g., the relay in the middle point of the line connecting S and D entails a reduction of $2^{\alpha-1} = 5.3$ in the power required to maintain a target SNR for $\alpha = 3.4$.

1.2 Wireless networks with infrastructure

In a wireless network with infrastructure a set of J user terminals $\{U_j\}_{j=1}^{J}$ communicates with any out of a set of J_{ap} infrastructure access points (APs) $\{U_j\}_{j=J+1}^{J_{\text{ap}}}$. The APs are interfaces to a reliable, usually wired, telecommunications infrastructure, implying that from a wireless networking perspective, the problem is to ensure that packets are reliably communicated to the infrastructure.

Since a wireless network comprises a collection of links $U_j \to U_i$ with $j \in [1, J]$ and $i \in [J+1, J+J_{\text{ap}}]$, its properties should reduce to those of the individual links. This is almost true for conventional wireless networks except for the necessity to separate individual transmissions. A general model of different separation techniques is to let the signal $x_j(t)$ transmitted by U_j be the product of the information bearing symbols $s_j(l)$ and a separation code $c_j(t)$, i.e.,

$$x_j(t) = \sum_{l=0}^{L-1} s_j(l) c_j(t - lT_s). \tag{1.17}$$

The signal received at any of the APs $\{U_i\}_{i=J+1}^{J+J_{\text{ap}}}$ is the superposition of the signals $\{x_j(t)\}_{j=1}^{J}$ transmitted by all terminals through channels h_{ij} plus AWGN noise

$$y_i(t) = \sum_{j=1}^{J} h_{ij}(t) x_j(t) = \sum_{j=1}^{J} x_j(t) + n(t). \tag{1.18}$$

Whereas in a point-to-point link the destination has to discern the symbols $s_S(l)$ from the AWGN noise $n(l)$ [cf. (1.4)] in a multiple access network the symbols of user U_j have to be separated from the noise $n(t)$ and the signals transmitted by other terminals.

It can be shown that a sufficient statistic to recover $\{\mathbf{s}_j\}_{j=1}^J$ is a bank of filters matched to the composite pulses $h_{ij}(t)c_j(t)$; see e.g., [69, Ch. 5]. We can thus define the discrete time received signals

$$y_{ij}(l) = \int_{lT_s}^{(l+1)T_s} z_i(t) \frac{[h_{ij}(t)c_j(t-lT_s)]^*}{\|h_{ij}(t)c_j(t-lT_s)\|}\, dt. \tag{1.19}$$

Proceeding as before we assume that the channels $h_{ij}(t)$ do not change during the duration of a symbol period T_s so that they can be factored out of the integral in (1.19). If we further define $\mathcal{C}_{kj}(l) := \int_{lT_s}^{(l+1)T_s} c_i^*(t-lT_s)c_j(t-lT_s)dt$ as the inner product between different user-separating codes we can write the equivalent discrete-time channel as

$$y_{ij}(l) = h_{ij}(l)s_j(l) + \sum_{k=1,k\neq j}^{J} h_{ik}(l)\mathcal{C}_{kj}(l)s_k(l) + n_{ij}(l). \tag{1.20}$$

Upon defining appropriate vectors and matrices (1.20) can be written in a more compact form. Define the aggregate transmitted symbols as the vector $\mathbf{s}(l) := [s_1(l), \ldots, s_J(l)]^T$, the vector received signal $\mathbf{y}_i(l) := [y_{i1}(l), \ldots, y_{iJ}(l)]^T$, and the noise vector $\mathbf{n}_i(l) := n_{i1}(l), \ldots, n_{iJ}(l)]^T$. Consider also the matrix of inner-products $\mathbf{C}(l)$ with (k,j)-th entry $\mathcal{C}_{kj}(l)$ and define the channel $\mathbf{H}_i(l) := \text{diag}[h_{i1}(l), \ldots, h_{iJ}(l)]$. Using these definitions we can write

$$\mathbf{y}_i(l) = \mathbf{C}(l)\mathbf{H}_i(l)\mathbf{s}(l) + \mathbf{n}(l). \tag{1.21}$$

Different multiple access techniques use different sets of codes $\{c_j(t)\}_{j=1}^J$ to separate transmitted signals. In time (T-) division multiple access (-DMA) the set of codes is simply a set of pulses with disjoint temporal support; in orthogonal frequency (OF-) DMA the codes are complex exponentials (frequency tones); and in code (C-) DMA are codes satisfying some orthogonality (or quasi-orthogonality) conditions. Each of these multiple access techniques finds application in different niches, but for the purpose of this thesis the fundamental difference is whether the codes are orthonormal or not.

When the codes $\{c_j(t)\}_{j=1}^{J}$ are orthonormal the correlations $\mathcal{C}_{ij}$ are null for any pair of disjoint codes and consequently the matrix $\mathbf{C}$ is the $J \times J$ identity matrix, i.e, $\mathbf{C} = \mathbf{I}$. In this case the communications do not interfere with each other and the multiple access channel becomes a simple collection of J separate channels of the form $U_j \rightarrow U_i$. When e.g., the channels are block fading, the relevant performance metric for each of these communications is the average SNR given by

$$\mathrm{SNR}_{ij} = \frac{P_j L(\mathbf{U}_i - \mathbf{U}_j)}{N_0}. \tag{1.22}$$

When the codes are not orthogonal, the optimal detector treats $\mathbf{s}(l)$ as a vector signal and performs joint detection of $\mathbf{s}(l)$ based on $\mathbf{y}_i(l)$. Even though optimal, this so called multiuser detector incurs computational complexity that grows exponentially with the number of terminals J. Thus, more often than not, a single user detector is used. In a single user detector, $s_j(l)$ is decoded by using $y_{ij}(l)$ only as per the signal model in (1.20) with the interference from users $U_k \neq U_j$ regarded as noise. In this sub-optimal scheme the pertinent figure of merit is the signal interference plus noise ratio (SINR) given by

$$\mathrm{SINR}_{ij} = \frac{P_j L(\mathbf{U}_j - \mathbf{AP})}{N_0 + \sum_{k=1, k\neq j}^{J} \mathcal{C}_{kj} P_k L(\mathbf{U}_i - \mathbf{AP})}. \tag{1.23}$$

Whether the SNR in (1.22) or the SINR in (1.23) is the pertinent figure of merit, the important point here is that a conventional wireless network can be modeled as set of links of the form $U_j \rightarrow U_i$. However, this neglects the fact that signals transmitted by user terminals are overheard by their peers. This naive observation leads naturally to consider collaborative networks.

1.3 Collaborative networks

The model of a conventional wireless network as described in the previous section is a collection of links between terminals $\{U_j\}_{j=1}^{J}$ and infrastructure APs $\{U_i\}_{j=J+1}^{J+J_{\mathrm{ap}}}$. Due to broadcast propagation however, the definition of a link in a wireless network is somewhat of a fatuous point, since the signal transmitted by U_j is overheard not only by the APs but also by other terminals. The idea of collaborative networks is to exploit the broadcast nature of the wireless channel by letting terminals relay packets for each other.

Thus, the signal model coincides with the one in (1.21), but instead of considering signal reception at the APs only, i.e., for $i \in [J+1, J+J_{ap}]$ we also consider signal reception at other user terminals. Repeating (1.21) for convenience we model signal reception as

$$\mathbf{y}_i(l) = \mathbf{C}(l)\mathbf{H}_i(l)\mathbf{s}(l) + \mathbf{n}(l). \tag{1.24}$$

with the latter expression considered for $i \in [1, J+J_{ap}]$, i.e., for user terminals and APs.

Given that (1.24) coincides with (1.21) the discussion following the latter in Section 1.2 is pertinent to collaborative networks as well. In particular, if the codes used by different terminals are orthogonal the probability R_{ij} of U_i decoding a packet from U_j is determined by the SNR in (1.22). If the codes are not orthogonal then the SINR in (1.23) might be of interest. However, it is important to note that not all terminals transmit at the same time and some random access considerations may have a role to play in this case, depending on the number of simultaneous transmissions and the burstiness of the traffic they generate..

The difference between conventional networks considered in Section 1.2 and collaborative networks considered here is that packets decoded at a user terminal $\{U_i\}_{i=1}^{J}$ are in fact intended for the APs $\{U_i\}_{i=J+1}^{J+J_{\text{ap}}}$. Thus, we need a mechanism to find multi-hop routes from user terminals to infrastructure APs. Since this problem is well studied in wired networks, why is there a need to "reinvent the wheel" for wireless collaborative networks?

In fact, as [27] correctly points out "we all have learned to draw a graph to depict a communication network" and not surprisingly most of the research in wireless networking concentrates on reducing the wireless network to a wired-like – i.e., graph – model. Consider, for example, the problem of multi-hop routing that we will study in Chapters 2 and 3. Many useful multi-hop routing algorithms adhere to the so called "disk routing models" which typically proceed in three stages: i) define a communication radius for each node; ii) draw the corresponding connectivity graph; and iii) utilize network optimization tools, e.g., shortest path routing, to find the optimal route. Most of the differences in multi-hop routing algorithms arise in the definition of the associated link metrics. These include path reliability, transmitted power, and mutual interference, to name a few; see e.g., [34, 90] and references therein. The problem with this approach is that since a link in a wireless network does not entail a tangible connection, its definition can be somewhat arbitrary. As

we increase the communication radius, link reliability decreases but should we define a link as any communication with reliability greater than say 70%?, or should the graph with 90% reliability be preferred? Truth is, there is no satisfactory answer to this question.

There is a growing consensus in the research community that there is a need to develop novel models to deal with wireless networks. As no definite model is available yet, one of the goals of this thesis is to contribute novel models towards a better understanding of wireless collaborative networks. To this end, our contributions are in the areas of routing, random access, and multiple access as we outline in the next section.

1.4 Roadmap

The research dealt with in this thesis contributes to the advancement of wireless collaborative networking (CN) aspiring to yield significant improvements in terms of capacity, coverage and error performance with respect to existing alternatives. These improvements stem from the diversity and pathloss reduction effected by user collaboration. On the one hand, collaboration provides alternative routes mitigating fading effects. On the other hand, multi-hop routes counteract the rapid decay in average received power. Accordingly, we can divide the contributions of this thesis in two interrelated thrusts:

[**T1**] **Multi-hop routing** Common approaches to designing muti-hop routing protocols start by building a connectivity graph of the wireless network. These approaches yield valuable results but do not fully exploit the benefits of the broadcast wireless channel. Our goal in this thrust is to develop routing protocols based on more accurate probabilistic models accounting for the broadcast nature of the wireless channel.

[**T2**] **Cooperation in ad-hoc, fixed and random multiple access networks.** User cooperation diversity has well appreciated merits in point-to-point links. Our goal here is to develop theory and methods to understand and exploit user cooperation in wireless networks.

Even though different in the principles they exploit, both thrusts commonly rely on what we purport as a shift in wireless networks: from competition towards collaboration.

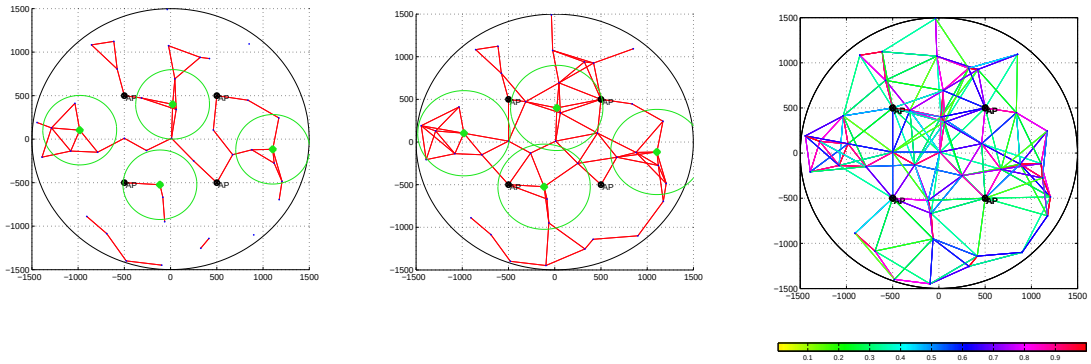

Figure 1.1: A network of 40 user terminals (blue and green) collaborate to route packets to any of the 4 infrastructure APs (black). The communication disk for representative nodes (green) and the resulting graph (red) are shown when the communication radii is 400 (left) and 500 (middle) meters, respectively. A better model is the reliability matrix $\mathbf{R}$ (left). The color-encoded (i, j)-th entry R_{ij} of $\mathbf{R}$ represents the probability that a packet transmitted from the j-th user U_j is correctly received by the i-th user U_i. (R_{ij} is generated according to the empirical distribution in [3].)

This shift we envision to offer a step towards achieving the broader goal of collaborative architectures which we aspire to impact ad-hoc networks, wireless sensor networks as well as next generation cellular and tactical networks.

1.4.1 Multi-hop routing in collaborative networks

As we mentioned before, the traditional approach to routing in multi-hop networks is to build a graph-theoretic model. If user nodes transmit over orthogonal channels it is plausible to assume that communication from U_j to U_i is feasible if and only if the average SNR exceeds a certain link reliability threshold. Recalling that P_j denote the transmit-power of U_j, N_0 the noise power at U_i, and $\mathbf{U}_j$ the position of U_j, we can express the condition for existence of a communication link as

$$\mathrm{SNR}_{ij} = \frac{P_j L(\mathbf{U}_j - \mathbf{U}_i)}{N_0} \geq \mathcal{T} \tag{1.25}$$

where $\mathcal{T}$ is a threshold ensuring that the link reliability is high enough.

We can now draw a network graph with arcs connecting pairs of nodes that satisfy (2.1) as depicted in Fig. 1.1. In order to find optimal routes from any user terminal $\{U_j\}_{j=1}^{J}$ to any of the APs $\{U_j\}_{j=J+1}^{J_{\rm ap}}$ network optimization tools, e.g., shortest path routing, are then utilized.

This so-called "disk model" effectively reduces wireless routing to routing over a wired network, thus inheriting a number of attractive properties. Of particular interest is the fact that an optimal route can be found in $O(J^2)$ steps using dynamic programming schemes implemented with the Bellman-Ford, Dijkstra, or Floyd-Warshall algorithms [8, Chap.5]. These routing algorithms can either be implemented at a central node, say any of the APs, or in a distributed manner relying on communication with one-hop neighbors only [9, Chap. 4]. However, Fig. 1.1 purposefully points to the arbitrary selection of the reliability threshold $\mathcal{T}$ since it reveals that different communication radii give rise to considerably different graphs. A small value of $\mathcal{T}$ yields a densely connected graph, and consequently routes with a small number of hops, but may lead to the use of unreliable links. A large $\mathcal{T}$ on the other hand, enforces the use of reliable links but the resultant graph entails more hops to reach the destination possibly winding up with a disconnected graph.

1.4.2 Routing in collaborative networks

Given the unsuitability of graph models to describe wireless networks, the natural step is to prescind of the graph model altogether and consider multi-hop routing as an optimization problem based on the reliability (i.e., the pairwise packet-success-probability) matrix $\mathbf{R}$ whose (i,j)-th entry R_{ij} represents the probability that a packet transmitted from the j-th user U_j is correctly received by the i-th user U_i [79]; see also Fig. 1.1.

While it is clear that $\mathbf{R}$ provides for a better model of the wireless network the issue is whether it is a more useful model. That is, can we obtain better insights by using $\mathbf{R}$? and can we design routing algorithms based on $\mathbf{R}$ that objectively outperform those based on graph models? This thesis contends that, indeed, using $\mathbf{R}$ in wireless multihop routing protocol design offers a powerful alternative because of:

- **Increased Rates.** When routing matrices are chosen to optimize rate metrics, the

use of $\mathbf{R}$ yields a larger set of achievable rates than those enabled by the disk model.

- **Reduced Complexity.** Optimization problems on a graph usually turn out to incur combinatorial complexity. Many optimization problems involving a matrix however, can be solved in polynomial time using convex optimization techniques [13]. The latter will turn out to be the case with our routing protocols which promise to be also attractive from a complexity perspective.

- **Novel routing criteria.** Many optimal routing criteria of practical interest are considered intractable since they entail graph optimization algorithms with combinatorial complexity. Our approach permits re-formulation of many such criteria and renders them tractable.

- **Distributed self-organizing implementations.** Even though convexity in optimization ensures manageable complexity, we still require $\mathbf{R}$ to be available at a central location. This entails: i) a large communication cost to collect $\mathbf{R}$ and percolate the optimal routing matrix; ii) possibly considerable delay to compute the optimal routes; and iii) lack of resilience to changes in $\mathbf{R}$, a problem particularly important in mobile scenarios. Many optimal routing problems can be solved by an iterative distributed algorithm whereby i) node U_j has access only to link reliability metrics for transmission to and from other nodes (the j-th row and column of $\mathbf{R}$); ii) U_j interchanges messages only with one-hop neighbors, defined as the set of terminals with non-zero probability of decoding U_j's packets; and iii) as time progresses U_j computes its optimal routing probabilities.

1.4.3 Diversity in random access channels

User cooperation was introduced as a diversity enabler for point-to-point wireless links whereby a pair of cooperating terminals share their respective information packets to create a virtual antenna array [49,51,96,97]. This way, each user is provided with two independent paths to the intended destination, the direct path and a second path relayed through the

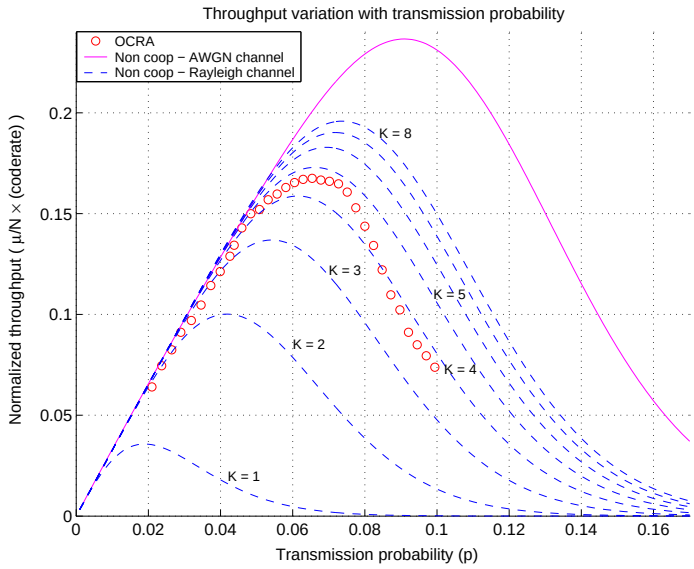

Figure 1.2: Cooperative random access (RA) can close the gap between the performance of RA over wireless Rayleigh fading and wired AWGN channels.

other user. It is not surprising that this virtual 2×1 multiple input - single output (MISO) channel provides second order diversity as its real i.e., non-virtual, counterpart [4, 108].

In random access (RA) networks, users access the channel at random [2]. With a sufficiently small channel access probability, the chance that two or more users decide to transmit in the same slot is small and we have packets successfully transmitted even when there is no coordination among users. Interestingly, the very random nature of RA dictates that in any time slot only a fraction of potential users is active, the others having their transmissions deferred. But since only a few out of the total number of transmitters are active at any given time, transmission hardware resources are inherently under-utilized in wireless RA networks. It is thus reasonable to expect that user cooperation can exploit these resources to gain a diversity advantage. In fact, intuition suggests that user cooperation appears to be a form of diversity well matched to RA [84, 85, 89, 127].

While the advantages of diversity at the physical layer are relatively well-known, the question remains as to how much we *stand to win* from cooperation at the medium access (MAC) layer. The answer is summarized in Fig. 1.2, where we plot throughput for a

Rayleigh fading channel, an additive white Gaussian noise (AWGN) channel and fading channels providing different orders of diversity. It comes as no surprise that throughput over the wireless (Rayleigh) channel is miserable, being almost an order of magnitude smaller than the throughput of the wired (AWGN) channel. This sizeable gap can be closed by diversity techniques, as hinted by the twofold increase observed with second-order diversity and the close-to-AWGN throughput enabled with eighth-order diversity. Eventually, as the diversity order keeps increasing the diversity-enriched channel approaches an AWGN channel.

While the potential gains are significant, a more relevant question is how much we *actually win* from cooperation at the MAC layer. We will show in this thesis that for "sufficiently large" networks the throughput of a cooperative RA network operating over a wireless channel approaches the throughput of an equivalent RA network operating over a wired channel. In other words, cooperation has the potential to *render wireless channels equivalent to wired ones*. How large is "sufficiently large" will be elaborated in Chapter 5. For now, it suffices to say that this claim is valid for moderately large networks. Fig. 1 depicts throughput for a cooperative RA network with 128 users from where we can verify that the throughput increases by an order of magnitude with respect to conventional non-cooperative RA protocols.

1.5 Published results

My Ph. D. work on cooperative communications and networking has resulted in the publication of 6 journals papers in the Institute of Electrical and Electronic Engineers (IEEE) Transactions on Wireless Communications [75, 76, 83], IEEE Transactions on Signal Processing [79], IEEE Transactions on Information Theory [85], and IEEE Journal on Selected Areas on Communications [88]. A tutorial paper featuring work in this thesis appeared in the European Signal Processing Society (EURASIP) Newsletter [77]. A second tutorial is scheduled to appear in the IEEE Signal Processing Magazine [78]. The work has also been disseminated at pertinent conferences where a total of 12 articles have been accepted for presentation [16, 73, 74, 80–82, 84, 86, 87, 89, 120, 127].

Chapter 2

Routing in collaborative networks

As discussed in Chapter 1, multi-hopping exhibits a significant potential to enable energy savings. Considering that received power decays exponentially with distance as $d^{-\alpha}$, with α between 3 and 4 – depending on the environment – the numbers are staggering. Splitting for instance a single hop in two hops can save as much as 10 dB in energy; and dividing a route in ten hops consumes in the order of a thousandth of the energy consumed by the original single hop [129]. Even though the former is a rough assessment, it is not difficult to appreciate that by reducing the average distance between communicating pairs of nodes multi-hopping secures significant power savings, if not the feasibility of the communication link itself.

The challenges to implement multi-hopping in wireless networks are many. Among the major ones is to find routes to the intended destinations that are optimal – in terms of for example, offered rate or power consumption – yet at the same time provide resilience against channel fades without requiring excessive levels of redundancy. The goal of this chapter is to introduce a general optimization framework for finding optimal stochastic routes in wireless multi-hop networks.

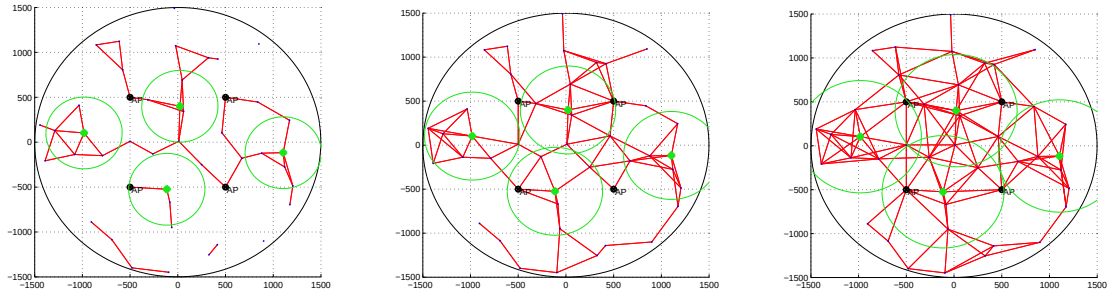

Figure 2.1: A network of 40 user terminals (blue and green) collaborate to route packets to any of the 4 infrastructure APs (black). To find routes to an AP, we can assume that communication between pairs of nodes is possible only when their distance is small enough. The communication disk for representative nodes (green) and the resulting graph (red) are shown when the communication radii is 400 (left), 500 (middle), and 600 (right) meters, respectively.

2.1 The routing problem in wireless

Consider the problem of collaborative routing to infrastructure in which a set of J wireless user terminals $\{U_j\}_{j=1}^{J}$ communicates with any out of a group of J_{ap} access points (AP) $\{U_j\}_{j=J+1}^{J+J_{\text{ap}}}$. An instrumental role in devising multi-hop routing algorithms is played the model adopted to describe what constitutes a link between two nodes in the wireless network. As [27] correctly points out "we all have learned to draw a graph to depict a communication network" and, not surprisingly, routing algorithms for wireless networks have evolved from the accumulated knowledge about these graph models. But since a link in a wireless network does not entail a tangible connection, associating links with arcs on a graph can be somewhat arbitrary. Nonetheless, if terminals transmit over orthogonal channels it is plausible to assume that communication from U_j to U_i is feasible if and only if the average signal-to-noise ratio (SNR) exceeds a certain link reliability threshold. Letting P_j denote the transmit-power of U_j, N_0 the noise power at U_i, and $\mathbf{x}_j$ the position of U_j, we can express the condition for existence of a communication link as [cf. (1.22)]

$$\text{SNR}_{ij} = \frac{P_j L(\mathbf{x}_j - \mathbf{x}_i)}{N_0} \geq \mathcal{T} \tag{2.1}$$

where $\mathcal{T}$ is a threshold ensuring that the link reliability is high enough, and $L(\mathbf{x}_j - \mathbf{x}_i)$ is a distance-dependent path loss coefficient. A usual model for $L(d)$ is an exponential path loss law for which $L(d) = d^{-\alpha}$.

We can now draw a network graph with arcs connecting pairs of nodes that satisfy (2.1) as depicted in Fig. 2.1. In order to find optimal routes from any user terminal $\{U_j\}_{j=1}^{J}$ to any of the APs $\{U_j\}_{j=J+1}^{J_{\text{ap}}}$ network optimization tools, e.g., shortest path routing, are then utilized. Models based on the SNR threshold are generically called "disk models" due to the fact that for exponential path loss laws U_j communicates only with nodes inside a disk of radius $[P_j/(\mathcal{T}N_0)]^{1/\alpha}$ centered at $\mathbf{x}_j$.

An approach to routing based on the disk model is to define U_j's optimal route to reach an AP as the one with the minimum number of hops in any of the graphs in Fig. 2.1. More generally we can associate a link metric M_{ij} with each arc $U_j \to U_i$ of the network graph in Fig. 2.1 and compute the shortest path route as the one minimizing the sum of the individual arcs along all possible routes. Choices of link metrics for multi-hop networks abound; see e.g., [90] and [103]. We can, for example, set $M_{ij} = P_j$, to obtain the routes of minimum power consumption [48].

The disk model effectively reduces wireless routing to routing over a wired network, thus inheriting a number of attractive properties. Of particular interest is the fact that an optimal route can be found in $O(J^2)$ steps using dynamic programming schemes implemented with the Bellman-Ford, Dijkstra, or Floyd-Warshall algorithms [8, Chap.5]. These routing algorithms can either be implemented at a central node, say any of the APs, or in a distributed manner relying on communication with one-hop neighbors only [9, Chap. 4]. However, Fig. 2.1 purposefully points to the arbitrary selection of the reliability threshold $\mathcal{T}$ since it reveals that different communications radii give rise to considerably different graphs. A small value of $\mathcal{T}$ yields a heavily connected graph, and consequently routes with a small number of hops, but may lead to the use of unreliable links. A large $\mathcal{T}$ on the other hand, enforces the use of reliable links but the resultant graph entails more hops to reach the destination possibly winding up with a disconnected graph.

It is also instructive to consider all the factors that (2.1) does not take into account.

These include reliability of isolated links due to fading and other factors, and interference from other links due to the broadcast nature of the wireless channel. For instance, rapid variations in the $U_j \to U_i$ link gain due to fading render it unlikely that the condition in (2.1) suffices to ensure a successful communication except for very low thresholds $\mathcal{T}$. The interference when terminals transmit over non-orthogonal channels is not considered either. The broadcast nature of the wireless channel implies that packets transmitted by U_j are not only received at the intended destination but overheard by other nodes in its neighborhood, thus bringing in question the link definition itself.

It is by now accepted that there is a need to develop novel models to deal with routing information in wireless multi-hop collaborative networks; see e.g., [35] and [36]. We first describe two representative approaches that alter the rules for defining the communication graph and the associated link metrics.

2.1.1 Accounting for interference

An attempt to account for interference is to modify (2.1) so that instead of requiring a sufficiently high SNR we require a sufficiently high signal-to-interference-plus-noise ratio (SINR). Assuming all transmitters are active all the time, the $U_j \to U_i$ link is added to the graph whenever [cf. (1.23)]

$$\mathrm{SINR}_{ij} = \frac{P_j L(\mathbf{x}_j - \mathbf{x}_i)}{N_0 + \gamma \sum_{k \neq i,j}^{J} P_k L(\mathbf{x}_k - \mathbf{x}_i)} \geq \mathcal{T} \tag{2.2}$$

with the term $\sum_{k \neq i,j}^{J} P_k L(\mathbf{x}_k - \mathbf{x}_i)$ denoting the power received at U_i from users different than U_j (listen-while-you-talk is infeasible on the same channel) and γ is the inverse of the processing gain of, e.g., a spread-spectrum system, scaling the effect of interference. Depending on the orthogonality between codes used during simultaneous transmission, γ is equal to 1 in a narrowband system, and is smaller than 1 in a broadband CDMA system.

Interestingly, the effect of interference is to deform the communication area of each node, which instead of a disk becomes dependent on the spatial distribution and transmission parameters (e.g., power) of nodes in its neighborhood. For the purposes of routing, while different from the one generated by (2.1), the link model in (2.2) still gives rise to a graph,

and the problem reduces to that of routing in a wired network.

All models discussed so far take a black / white (i.e., link / no link) approach to the modeling of individual links thus ignoring the (possibly significant) "gray areas" with intermediate link reliability [59]. This limitation motivates the approaches we will focus on in the rest of the chapter.

2.1.2 Link reliability as a metric

A first attempt to account for link reliability is to consider arc metrics given as functions of it. Ideally, we want a metric taking small non-negative values for links with large SNR, increasing as the SNR decreases, and eventually growing to infinity as the SNR goes to zero – amounting to absence of a link in the graph. The inverse SNR metric, i.e., $M_{ij} = 1/\text{SNR}_{ij}$, exhibits such a behavior thus mitigating the problem of using unreliable hops mentioned in the previous section [128].

Since the (average) SNR measures link reliability only to the extent that it determines the error probability of the given link, it is more natural to consider arc metrics depending on the link packet success probability R_{ij}. Again, we seek an inverse relation between an arc metric and link reliability as the one provided by $M_{ij} = 1/R_{ij}$. The cost per hop in this case ranges from one for a perfectly reliable link to infinity for a link with zero reliability.

The metric $1/R_{ij}$ has an interesting interpretation. In a link with reliability R_{ij}, out of x transmitted packets $R_{ij}x$ are correctly received, meaning that to have one packet correctly received ($x = 1$) an average of $1/R_{ij}$ packets must be transmitted from U_j to U_i. Furthermore, since the decoding probabilities do not depend on the history of the packet across hops the average number of times a packet is transmitted from its source to an AP terminal is the sum of $1/R_{ij}$ over the links belonging to the route used. Thus, the shortest path route when using the metric $M_{ij} = 1/R_{ij}$ is the one that minimizes the average number of times a given packet is transmitted [10, 22].

If we neglect queuing and processing delays (possible in a lightly loaded network), these routes also minimize the overall transmission delay [80]. Consequently, use of the $1/R_{ij}$ metric is justified for non-real time applications, for example file transfers, in which average

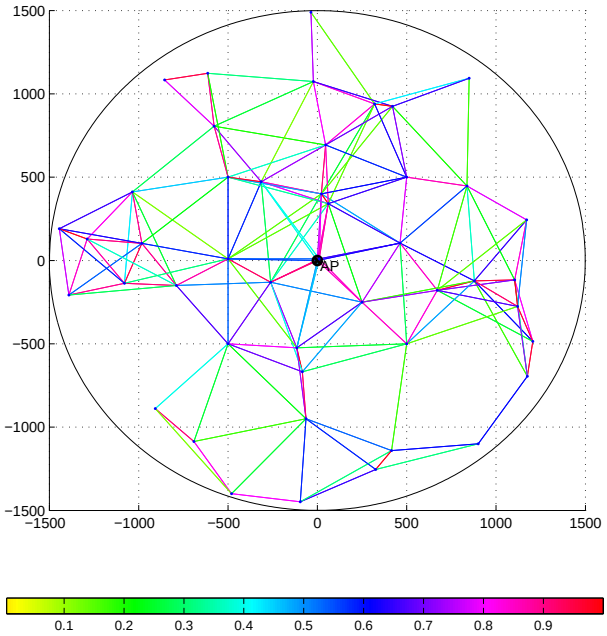

Figure 2.2: Schematic representation of the reliability matrix $\mathbf{R}$. The color-encoded (i, j)-th entry R_{ij} of $\mathbf{R}$ represents the probability that a packet transmitted from the j-th user U_j is correctly received by the i-th user U_i. The reliability matrix $\mathbf{R}$ is a better-suited model than the graphs in Fig. 2.1. (R_{ij} is generated according to the empirical distribution in [22]; only arcs with $R_{ij} > 0.3$ are shown.)

delay is of interest. For real time applications, for example voice and/or video conferencing, one is interested in minimizing worst case delays and the route minimizing the average delay is not necessarily optimal.

2.2 Stochastic routing

Given the unsuitability of graph models to describe wireless networks, the natural step is to prescind of the graph model altogether and consider multi-hop routing as an optimization problem based on the reliability (i.e., the pairwise packet-success-probability) matrix $\mathbf{R}$ whose (i, j)-th entry R_{ij} represents the probability that a packet transmitted from the j-th user U_j is correctly received by the i-th user U_i [79]; see also Fig. 2.2 and [3].

Specifically, consider the same setup discussed in Section 2.1 with $J_{\rm ap} = 1$ (extensions are straightforward and discussed in Section 2.4). Consider thus, a wireless network with $J+1$ user nodes $\{U_j\}_{j=1}^{J+1}$ in which the first J users $\{U_j\}_{j=1}^{J}$ participate in routing packets to the destination $D \equiv U_{J+1}$. The physical and medium access layers are such that if a packet is transmitted by U_j it is correctly *received* by U_i with probability R_{ij} that we arrange in the matrix $\mathbf{R}$. Note that in the presence of fading the probabilities R_{ij} are averaged over all fading states. We first consider a *per-session* model of routing in which a user node establishing a session is confronted with the routing decisions of its peers that determine the entries R_{ij} of $\mathbf{R}$. Supposing that the probabilities in $\mathbf{R}$ remain invariant over the duration of a session, our goal is to find a stochastic routing strategy that is optimal in a suitable sense. Note that this model is also applicable in a low traffic scenario, where at any time there is only one packet in the network.

Let $e_j(n)$ indicate the binary (0/1) event that the packet is at U_j at time n whose probability we denote by $p_j(n) := \Pr\{e_j(n) = 1\}$. Correspondingly, we define the vectors $\mathbf{e}(n) := [e_1(n), \ldots, e_{J+1}(n)]^T$ and $\mathbf{p}(n) := [p_1(n), \ldots, p_{J+1}(n)]^T$, where ^{T} denotes transposition. If the packet is generated at a known source U_s for some $s \in [1, J]$ we have that $p_s(0) = 1$. In general, the packets are generated at a random source with initial distribution $\mathbf{p}(0)$.

Routing is carried on according to a matrix $\mathbf{T}$ whose (i, j)-th entry T_{ij} is the probability that U_j decides to *transmit* (i.e., route) the packet to U_i. If U_j receives the packet at a certain time n, i.e., if $e_j(n) = 1$, U_j will select a random candidate destination from the set $\{U_i\}_{i=1}^{J+1}$ such that U_i is chosen with probability T_{ij}. If the transmitted packet is correctly decoded by U_i we have that $e_i(n+1) = 1$; otherwise, the packet is kept by U_j, i.e., $e_j(n + 1) = 1$, and the random selection and transmission process is repeated. To describe the stochastic percolation of the packet throughout the network we define the matrix $\mathbf{K}$ with (i, j)-th entry $K_{ij} := \Pr\{e_i(n + 1)|e_j(n)\}$ denoting the probability that the packet moves from U_j to U_i between times n and $n + 1$. Note that $\mathbf{T}$ and $\mathbf{K}$ are related through $\mathbf{R}$. Indeed, for $i \neq j$ the packet moves from U_j to U_i if and only if it is routed through U_i

and is correctly decoded; since these two events are independent we have

$$K_{ij} = T_{ij}R_{ij} \quad \text{for } i \neq j. \tag{2.3}$$

Because $\mathbf{K}$ and $\mathbf{T}$ are stochastic matrices, columns must sum up to 1 implying that $\mathbf{K}^T\mathbf{1} = \mathbf{1}$ and $\mathbf{T}^T\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes the all-one column vector. These two constraints and (3.1) imply that since $\mathbf{R}$ is prescribed by the physical layer, $\mathbf{K}$ is uniquely determined by $\mathbf{T}$ (but not vice versa).

Since the $(J+1)$-st user is the destination it will not route the packet, from which we infer that $T_{i(J+1)} = 0$, $\forall i \in [1, J]$; and after taking (3.1) into account we arrive at $K_{i(J+1)} = 0$, $\forall i \in [1, J]$. Arguing similarly, it follows that $R_{(J+1)(J+1)} = T_{(J+1)(J+1)} = K_{(J+1)(J+1)} = 1$. Summing up, with properly defined $\mathbf{k}_1 \in \mathbb{R}^J$ and $\mathbf{K}_0 \in \mathbb{R}^{J\times J}$ we can partition $\mathbf{K}$ as

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_0 & \mathbf{0} \\ \mathbf{k}_1^T & 1 \end{pmatrix}_{(J+1)\times(J+1)}, \tag{2.4}$$

where $\mathbf{0}$ denotes the all-zero column vector. Let $\mathbf{c}_{J+1} := [0, \ldots, 0, 1]$ denote the $(J+1)$-st vector in the canonical basis of $\mathbb{R}^{J+1}$. It follows easily by direct substitution that (2.4) holds if and only if $\mathbf{K}\mathbf{c}_{J+1} = \mathbf{c}_{J+1}$, i.e., if and only if $\mathbf{c}_{J+1}$ is an eigenvector of $\mathbf{K}$ associated with the eigenvalue 1.

For future reference, we define the set of transmit probability matrices in $\mathbb{R}^{(J+1)^2}$ as

$$\mathcal{T} = \{\mathbf{T} \in \mathbb{R}^{(J+1)^2} \ : \ \mathbf{T}^T\mathbf{1} = \mathbf{1}, T_{ij} \geq 0, \forall i, j\}. \tag{2.5}$$

The constraints on $\mathbf{K}$ can be written as $\mathbf{K} \in \mathcal{K}$ with

$$\mathcal{K} = \{\mathbf{K} \in \mathcal{T} : K_{ij} = T_{ij}R_{ij}, \text{for } i \neq j, \mathbf{T} \in \mathcal{T}; \ \mathbf{K}\mathbf{c}_{J+1} = \mathbf{c}_{J+1}\}. \tag{2.6}$$

Note that the set $\mathcal{K}$ is a convex polyhedron in $\mathbb{R}^{(J+1)^2}$.

We can characterize the evolution of $\mathbf{p}(n)$ in terms of $\mathbf{K}$. Indeed, note that due to the law of total probability $p_i(n) = \sum_{j=1}^n \Pr\{e_i(n)|e_j(n-1)\}p_j(n-1) = \sum_{j=1}^n K_{ij}p_j(n-1)$, that we can write in vector-matrix form as

$$\mathbf{p}(n) = \mathbf{K}\mathbf{p}(n-1) = \mathbf{K}^n\mathbf{p}(0). \tag{2.7}$$

That is, $\mathbf{p}(n)$ represents the probability evolution of a Markov chain characterized by $\mathbf{K}$ in which the j-th state represents the presence of the packet at user node U_j.

2.2.1 Deliverability

A basic requirement for the routing matrix $\mathbf{T}$ is to ensure that packets are eventually delivered to the destination $D \equiv U_{J+1}$, i.e.,

$$\lim_{n\to\infty} \mathbf{p}(n) = \mathbf{c}_{J+1}, \tag{2.8}$$

Since it is meaningful to focus on routing matrices that, at least, satisfy (2.8), we introduce the following definition.

Definition 1 *A routing matrix* $\mathbf{T}$ *ensures deliverability if and only if (2.8) holds for any initial distribution* $\mathbf{p}(0)$.

Building on (2.7), it is possible to find conditions to ensure deliverability of an SR matrix as we describe in the following theorem.

Theorem 1 *The following statements are equivalent:*

(i) The routing matrix $\mathbf{T}$ *ensures deliverability.*

(ii) Matrix $\mathbf{K}$ *describes the probability evolution of an absorbing Markov chain whose unique absorbing state is* $J+1$.

(iii) The spectral radius of $\mathbf{K}_0$ *is strictly smaller than one, i.e., with* $\mathrm{eig}(\mathbf{K}_0)$ *denoting the set of eigenvalues of* $\mathbf{K}_0$ *we have* $\rho(\mathbf{K}_0) := \max|\mathrm{eig}(\mathbf{K}_0)| < 1$.

(iv) The matrix $\mathbf{K}_0$ *and the vector* $\mathbf{k}_1$ *in (2.4) satisfy* $\mathbf{k}_1^T(\mathbf{I}-\mathbf{K}_0)^{-1} = \mathbf{1}^T$.

Proof: Using induction we can easily show that the n^{th} power of $\mathbf{K}$ can be written as [cf. (2.4)]

$$\mathbf{K}^n = \begin{pmatrix} \mathbf{K}_0^n & \mathbf{0} \\ \mathbf{k}_1^T \sum_{k=0}^{n-1} \mathbf{K}_0^k & 1 \end{pmatrix}. \tag{2.9}$$

Upon defining $\mathbf{p}_D(n) := [p_1(n), \ldots, p_J(n)]^T$ containing the probabilities of finding the packet at any node other than the destination, it follows from (2.7) and (2.9) that

$$\mathbf{p}_D(n) = \mathbf{K}_0^n \mathbf{p}_D(0). \tag{2.10}$$

On the other hand, note that (2.8) is true if and only if $\lim_{n\to\infty} \mathbf{p}_D(n) = \mathbf{0}$.

To go from (i) to (ii) note that since for any $\mathbf{K} \in \mathcal{K}$, $\mathbf{K}\mathbf{c}_{J+1} = \mathbf{c}_{J+1}$, $J+1$ is by definition an absorbing state of the Markov chain defined by $\mathbf{K}$. If $j \neq J+1$ is another absorbing state then $\mathbf{K}\mathbf{c}_j = \mathbf{c}_j$ and for $\mathbf{p}(0) = \mathbf{c}_j$ we have that $\mathbf{K}^n\mathbf{p}(0) = \mathbf{c}_j$ for every n; thus, $\lim_{n\to\infty} \mathbf{p}(n) = \mathbf{c}_j \neq \mathbf{c}_{J+1}$. This is a contradiction if $\mathbf{T}$ ensures deliverability and consequently $J+1$ is the unique absorbing state.

If (iii) is not true, then $\lim_{n\to\infty} \mathbf{K}_0^n \neq \mathbf{0}$. Hence, there exists a vector $\mathbf{p}(0) \neq \mathbf{0}$ for which $\lim_{n\to\infty} \mathbf{K}_0^n\mathbf{p}(0) \neq \mathbf{0}$ implying that $J+1$ is not a unique absorbing state. Thus, (ii) implies (iii).

That (iii) implies (iv) follows after noting that since $\mathbf{1}^T\mathbf{K} = \mathbf{1}^T$, we have that $\mathbf{1}^T\mathbf{K}^n = \mathbf{1}^T$ and asymptotically $\lim_{n\to\infty} \mathbf{1}^T\mathbf{K}^n = \mathbf{1}^T$. But since (iii) also implies that $\lim_{n\to\infty} \mathbf{K}_0^n = \mathbf{0}$, we must have $\lim_{n\to\infty} \mathbf{k}_1^T \sum_{k=0}^{n-1} \mathbf{K}_0^k = \mathbf{1}^T$. To obtain (iv), note that the geometric series is such that $\sum_{k=0}^{\infty} \mathbf{K}_0^n = (\mathbf{I} - \mathbf{K}_0)^{-1}$.

Finally, if (iv) is true then we can use the fact that $\mathbf{K}_0^n$ is a stochastic matrix, i.e., $(\mathbf{K}_0^n)^T\mathbf{1} = \mathbf{1}$) to conclude that $\lim_{n\to\infty} \mathbf{K}^n = [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{1}]^T$ implying that (i) is true. □

Theorem 1 gives necessary and sufficient conditions for an SR matrix to have guaranteed deliverability. None of these conditions is difficult to achieve and, in general, simple routing algorithms, e.g., a random walk through the network with $T_{ij} = 1/J$, can ensure deliverability. A more interesting problem is how to obtain a matrix which guarantees that the limit in (2.8) is practically achieved with n as small as possible. This motivates different routing algorithms that we can obtain from (2.7) and analyze next.

2.2.2 Fastest convergence rate routing

The rate of convergence can be either measured on average or for the worst possible initial distribution $\mathbf{p}(0)$. These metrics lead to different criteria for optimal routing. Optimal routing in an average sense will be considered in Section 2.2.3. What we expect from an optimal routing matrix $\mathbf{T}$ is for the convergence rate in (2.8) to be as fast as possible. The distance – in some sense – between $\mathbf{p}(n)$ and $\mathbf{c}_{J+1}$ can be measured by the p-norm $\|\mathbf{p}(n) - \mathbf{c}_{J+1}\|_p$ which is to be compared with the original distance $\|\mathbf{p}(0) - \mathbf{c}_{J+1}\|_p$ leading

to the following expression for the worst-case convergence rate:

$$\xi_p = \sup_{\mathbf{p}(0)\neq\mathbf{c}_{J+1}} \lim_{n\to\infty} \left(\frac{\|\mathbf{p}(n) - \mathbf{c}_{J+1}\|_p}{\|\mathbf{p}(0) - \mathbf{c}_{J+1}\|_p} \right)^{1/n}. \tag{2.11}$$

This cannot be computed in closed-form for any p-norm. For $p = 2$, corresponding to the Euclidean norm, the argument in (2.11) is maximized by the eigenvector associated with the second largest eigenvalue of $\mathbf{K}$. A meaningful routing algorithm is thus to look for the matrix $\mathbf{K} \in \mathcal{K}$ such that

$$\min_{\mathbf{K}\in\mathcal{K}} |\text{eig}_2(\mathbf{K})| = \min_{\mathbf{K}\in\mathcal{K}} \max |\text{eig}(\mathbf{K}_0)| = \min_{\mathbf{K}\in\mathcal{K}} \rho(\mathbf{K}_0), \tag{2.12}$$

where $\text{eig}_2(\mathbf{K})$ denotes the second largest eigenvalue of $\mathbf{K}$ and $\text{eig}(\mathbf{K}_0)$ the set of eigenvalues of $\mathbf{K}_0$. In establishing the first equality in (2.12) we used that all the eigenvalues of $\mathbf{K}_0$ are eigenvalues of $\mathbf{K}$ [cf. (2.4)]; in fact, $\text{eig}(\mathbf{K}) = \text{eig}(\mathbf{K}_0) \cup \{1\}$. The second equality follows from the definition of spectral radius.

Unfortunately, minimizing the spectral radius of a non-symmetric matrix is a notoriously difficult problem, intractable except for small-medium values of J [13]. This motivates an alternative measure of convergence rate based on the vector $\mathbf{p}_D(n) := [p_1(n), \ldots, p_J(n)]^T$ containing the probabilities that the packet is at a certain node other than the destination. The norm of $\mathbf{p}_D(n)$ measures the probability of the packet *not* being delivered at time n. This suggests the metric

$$\zeta_p = \max_{\mathbf{p}_D(n)} \frac{\|\mathbf{p}_D(n+1)\|_p}{\|\mathbf{p}_D(n)\|_p}, \tag{2.13}$$

which amounts to the worst-case one-step relative reduction of the vector $\mathbf{p}_D(n)$ which we want converging to zero [cf. (2.8)]. Similarly to ξ_p, we can define optimal routing in terms of minimizing ζ_p.

If we further recall that $\mathbf{p}_D(n+1) = \mathbf{K}_0\mathbf{p}_D(n)$, another class of optimal SRPs stemming from (2.13) can be designed to achieve

$$\min_{\mathbf{K}\in\mathcal{K}} \max_{\mathbf{p}_D(n)} \frac{\|\mathbf{K}_0\mathbf{p}_D(n)\|_p}{\|\mathbf{p}_D(n)\|_p} = \min_{\mathbf{K}\in\mathcal{K}} \|\mathbf{K}_0\|_p, \tag{2.14}$$

where the equality follows from the definition of the p-norm of a matrix. Different from (2.12), the optimization in (2.14) is a convex problem for all p since: i) due to the triangle

inequality, norms are convex functions of their arguments; and ii) the set $\mathcal{K}$ is a convex polyhedron [cf. (2.6)]. For the usual norms, $p = 1, 2, \infty$, solving (2.14) is either a simple linear program (LP) for $p = 1, \infty$, or, a semi-definite program (SDP) for $p = 2$ [13].

In general, (2.12) and (2.14) are optimized by different matrices $\mathbf{T}$, and the pertinent comparisons are discussed in the following remark.

Remark 1 Requiring the solution of convex optimization problems – indeed, canonical optimization problems – (2.14) is tractable for networks with even a large number of users J; whereas (2.12) is only tractable for small-to-medium scale networks. On the other hand, (2.12) is more meaningful than (2.14), since the former compares the asymptotic behavior with the initial state while the latter compares two consecutive states. In practical protocol designs, (2.14) can be viewed as a tractable approximation to (2.12).

2.2.3 Minimum expected delay routing

An alternative approach to optimal routing is to consider the packet delivery time measured by the number of hops, and look for the matrix $\mathbf{T}$ that minimizes the average packet delay. Packet delay is simply the time n at which the packet is received by $D \equiv U_{J+1}$ and is given by:

$$\delta = \min\{n \ : \ e_{J+1}(n) = 1\} = \sum_{n=0}^{\infty} [1 - e_{J+1}(n)] \tag{2.15}$$

where the second equality is true since $1 - e_{J+1}(n) = 1$ if $n < \delta$ and $1 - e_{J+1}(n) = 0$ for $n \geq \delta$; we thus have δ terms equal to 1 in the summation in (2.15). Starting from (2.15), the expected delay can be computed as we describe in the following theorem.

Theorem 2 *For a routing matrix ensuring deliverability, the expected delay is given by*

$$\bar{\delta} := \mathrm{E}(\delta) = \mathbf{1}^T (\mathbf{I} - \mathbf{K}_0)^{-1} \mathbf{p}_D(0), \tag{2.16}$$

where $\mathbf{p}_D(0) := [p_1(0), \ldots, p_J(0)]^T$ *is the initial distribution for the first* J *users.*

Proof: Taking expected value in (2.15), using the linearity of the expected value operator and noting that $p_D(n) = \mathrm{E}[e_D(n)]$, we obtain (recall that $\sum_{j=1}^{J+1} p_j(n) = 1$)

$$\bar{\delta} = \sum_{n=1}^{\infty}[1 - p_{J+1}(n)] = \sum_{n=1}^{\infty}\sum_{j=1}^{J} p_j(n), \tag{2.17}$$

Writing the innermost summation as $\mathbf{1}^T\mathbf{p}_D(n)$ and recalling that $\mathbf{p}_D(n) = \mathbf{K}_0^n\mathbf{p}(0)$, we obtain

$$\bar{\delta} = \sum_{n=1}^{\infty}\mathbf{1}^T\mathbf{p}_D(n) = \sum_{n=1}^{\infty}\mathbf{1}^T\mathbf{K}_0^n\mathbf{p}(0) = \mathbf{1}^T\left(\sum_{n=1}^{\infty}\mathbf{K}_0^n\right)\mathbf{p}(0). \tag{2.18}$$

For routing matrices that ensure deliverability, Theorem 1 states that the spectral radius of $\mathbf{K}_0$ is $\rho(\mathbf{K}_0) < 1$. Consequently, the matrix geometric series in (2.18) is convergent with $\sum_{n=0}^{\infty}\mathbf{K}_0^n = (\mathbf{I} - \mathbf{K}_0)^{-1}$. Substituting this into (2.18), (2.16) follows readily. □

The expected delay $\bar{\delta}$ is a function of the routing matrix $\mathbf{K}$ *and* the initial distribution $\mathbf{p}_D(0)$. Using the result in Theorem 2, we can find the matrix that minimizes the expected delay as the argument solving the optimization problem

$$\mathbf{K}^*[\mathbf{p}_D(0)] = \arg\min_{\mathbf{K}\in\mathcal{K}}\bar{\delta} = \arg\min_{\mathbf{K}\in\mathcal{K}}\mathbf{1}^T(\mathbf{I} - \mathbf{K}_0)^{-1}\mathbf{p}_D(0). \tag{2.19}$$

Conceptually, (2.19) appears difficult to solve. Interestingly, it turns out that (2.19) is equivalent to a shortest path routing algorithm as we establish in the ensuing theorem.

Theorem 3 *Define the expected delay vector* $\bar{\boldsymbol{\delta}} := [\bar{\delta}_1, \ldots, \bar{\delta}_J] := \mathbf{1}^T(\mathbf{I} - \mathbf{K}_0)^{-1}$ *in which* $\bar{\delta}_j$ *is the expected delay when the packet starts at* U_j*, i.e., when* $\mathbf{p}(0) = \mathbf{c}_j$*; and let* $\bar{\delta}_{J+1} = 0$*. If there exists a matrix* $\mathbf{K}$ *ensuring deliverability, there exists a matrix* $\mathbf{K}^\dagger \in \mathcal{K}$ *such that*

$$\bar{\delta}_j = \min_i\left\{\frac{1}{R_{ij}} + \bar{\delta}_i\right\}, \quad \bar{\delta}_{J+1} = 0, \tag{2.20}$$

which minimizes the expected delay for any initial distribution, i.e., $\mathbf{K}^*[\mathbf{p}_D(0)] = \mathbf{K}^\dagger$ *for any* $\mathbf{p}_D(0)$ *and its corresponding* $\mathbf{K}^*[\mathbf{p}_D(0)]$ *as in (2.19).*

Proof: See Appendix A. □

Characterizing the solution as in (2.20) indicates that $\mathbf{K}_0^*$ in (2.19) can be found as the shortest path route (SPR) in a fully connected graph with the arc between U_i and U_j having weight $1/R_{ij}$. Indeed, let $\mathbf{i} := (i_1, \ldots, i_k)$ with $k \in [2, J+1]$, $i_1 = j$ and $i_k = J+1$ be

Algorithm 1 Min. expected delay routing (Dijkstra version)

Require: The packet success probability matrix $\mathbf{R}$

Ensure: The routing matrix $\mathbf{T}$

$\bar{\delta}_j = 1/R_{(J+1)j}$, for $j \in [1, J]$

$\mathcal{U} = \{U_j\}_{j=1}^{J}$

while $\mathcal{U} \neq \emptyset$ **do**

  $j^* = \arg\min_{j:U_j \in \mathcal{U}} \bar{\delta}_j$

  $\mathcal{U} = \mathcal{U} - \{U_{j^*}\}$

  for all $i : U_i \in \mathcal{U}$ **do**

    if $1/R_{ij^*} + \bar{\delta}_j^* < \bar{\delta}_i$ **then**

      $\bar{\delta}_i = 1/R_{ij^*} + \bar{\delta}_j^*$,

      $T_{ij^*} = 1$; $T_{ij} = 0$ for $j \neq j^*$

    end if

  end for

end while

an arbitrary sequence starting at U_j and finishing at U_{J+1}. Proceeding recursively, we find that (2.20) is equivalent to

$$\bar{\delta}_j = \min_{\mathbf{i}} \left\{ \sum_{l=1}^{\#(\mathbf{i})-1} \frac{1}{R_{i_l i_{l+1}}} \right\}, \tag{2.21}$$

where $\#(\mathbf{i})$ denotes the cardinality of $\mathbf{i}$. By definition, (2.21) is the SPR between j and $J+1$ among all the possible routes $\mathbf{i}$. In fact, the relation in (2.20) is Bellman's principle of optimality, which is known to characterize the shortest path route [8, Chap.5]. This implies that the solution to minimum expected delay routing can be found in $O(J^2)$ steps using dynamic programming tools, e.g., Bellman-Ford, Dijkstra, or Floyd-Warshall algorithms; see e.g., [8, Chap.5].

Also important, and contrary to what (2.19) suggested, minimum expected delay routing does not depend on the initial distribution. The average delays $\bar{\delta}[\mathbf{p}(0)]$ for different initial distributions $\mathbf{p}(0)$ are different, but there exists a matrix that minimizes $\bar{\delta}[\mathbf{p}(0)]$ for *all* $\mathbf{p}(0)$.

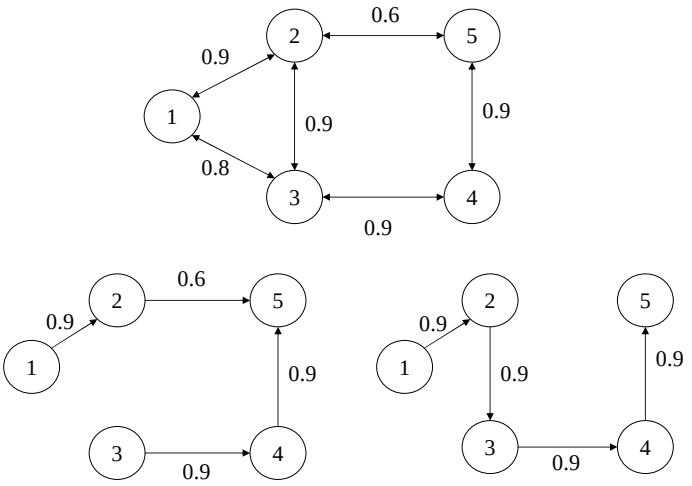

Figure 2.3: For a simple connectivity graph (top) the minimum expected delay routing algorithm in (2.19) tends to select short routes (left), while fastest convergence rate routing as per (2.12) selects longer routes with more reliable hops (right).

Among other optimization problems, such a matrix is the solution of the problem

$$\mathbf{K}^* = \arg\min_{\mathbf{K}\in\mathcal{K}} \mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-1}\mathbf{1} \tag{2.22}$$

obtained by making $\mathbf{p}_D(0) = \mathbf{1}/J$ in (2.19). Note that for a given $\mathbf{p}(0)$ there might exist alternative solutions to (2.19), but none will outperform $\mathbf{K}^*$ in (2.22). The matrix $\mathbf{K}^*$ in (2.22) can be obtained using Algorithm 1, a fact that we will later exploit in making pertinent comparisons between different routing algorithms.

2.2.4 Numerical examples and simulations

The fastest convergence rate SR algorithm in (2.12) maximizes the packet delivery probability for a given, sufficiently large, time index n. On the other hand, minimum expected delay routing as per (2.19) minimizes the expected time elapsed until packet delivery. The subtle differences between these two approaches are exemplified in Figs. 2.3 and 2.4.

The resulting routing matrices for minimum expected delay and fastest convergence rate routing are shown in Fig. 2.3. We can see that the former algorithm tends to select short routes sometimes containing unreliable hops (left) as verified by the link $U_2 \to U_5$ used to

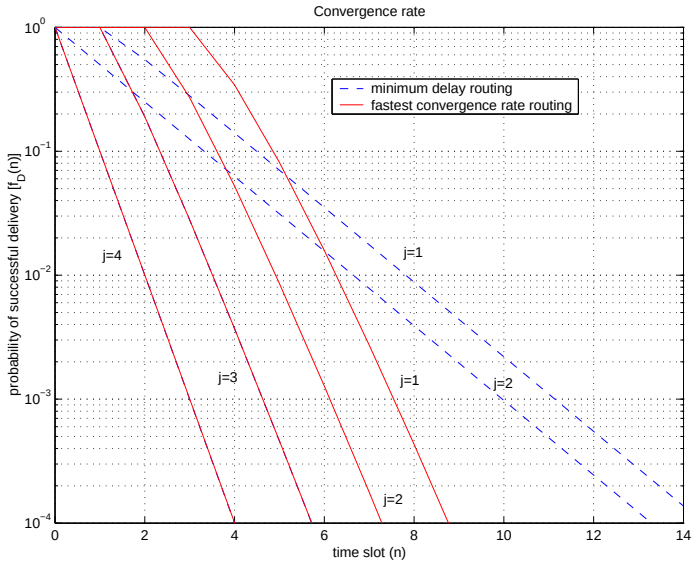

Figure 2.4: Convergence rate for the network in Fig. 2.3. For a fixed time delay fastest convergence rate routing yields a smaller packet error probability.

route U_1 and U_2's packets. Whereas, the latter uses longer routes but tends to use more reliable hops (right), as we can see from the use of the $U_2 \rightarrow U_3$ link to route U_1 and U_2's traffic. This is a manifestation of the different optimization criteria. The expected delay for routing U_2's packets is 1.67 for minimum expected delay routing and 3.33 for fastest convergence rate routing. The difference in convergence rate is shown in Fig. 2.4. To achieve a packet error probability of $1 - p_D(n) = 10^{-4}$, U_2's delay is 7.2 for fastest convergence rate routing and 13.1 for minimum expected delay routing.

Similar conclusions are reached for the more realistic example in Fig. 2.5 representing a randomly generated network with 20 nodes. In this figure, we depict the connectivity graph as well as the result of the minimum expected delay, fastest convergence rate, and minimum 2-norm SRP obtained from (2.14) with $p = 2$. Here it is also true that minimum expected delay prefers shorter routes, while fastest convergence rate prefers longer routes containing more reliable hops. Minimum 2-norm routing is the only algorithm considered that yields routing matrices implying non-deterministic routing, i.e., having $T_{ij} \neq 1, 0$ for some i, j.

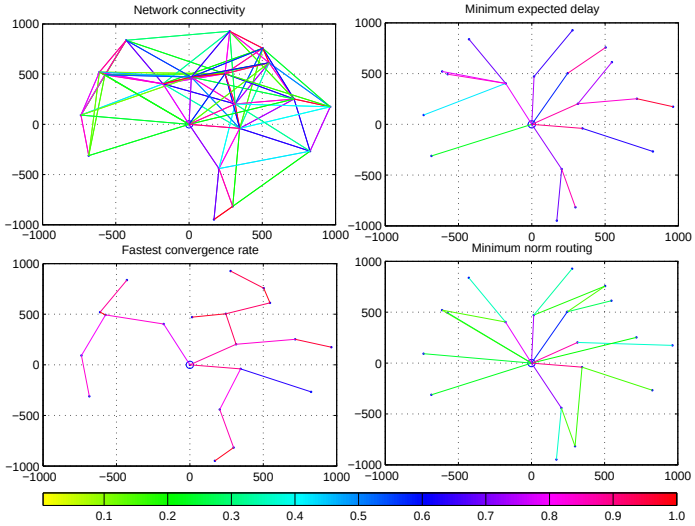

Figure 2.5: A randomly generated network with 20 nodes, the color scale represents the elements of the matrix $\mathbf{K}$. Note how fastest convergence rate routing selects routes with large values of K_{ij}.

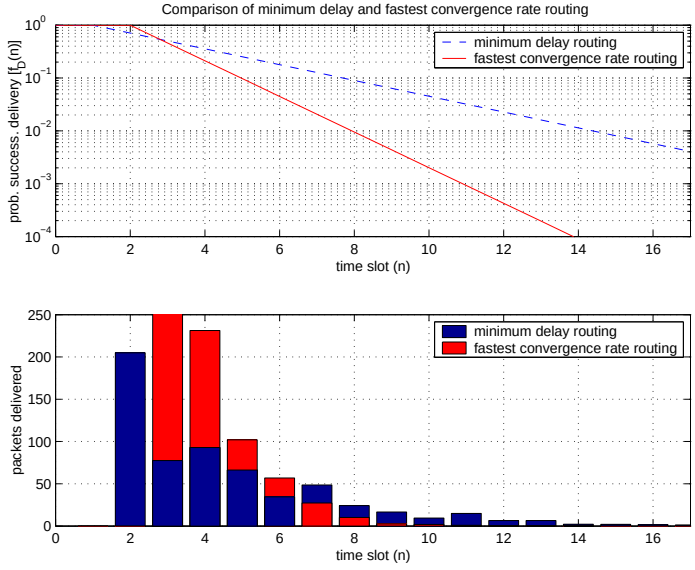

Figure 2.6: Convergence rate of the least favored user for the network in Fig. 2.5 (top) and histogram of packet delivery times for a randomly chosen user (bottom). Fastest convergence rate routing is favored for time sensitive traffic.

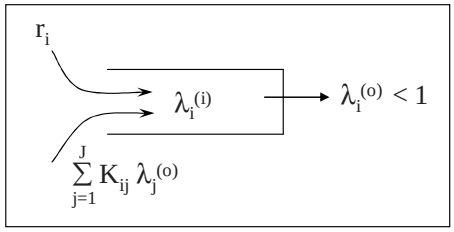

Figure 2.7: Queue balance equations.

For real time delay-sensitive applications, e.g., audio and/or video conferencing, fastest convergence routing is a better alternative. This is corroborated by Fig. 2.6 (top) showing the convergence rate for the network in Fig. 2.5. For a delay of 14 hops, fastest convergence rate routing yields a packet error probability of 10^{-4} for the least favored user; for the same delay, minimum expected delay routing achieves a packet error probability of 10^{-2}. For delay-tolerant applications, e.g., file transfers, the average delay metric is better suited since to deliver a large number of packets, the total number of required hops is significantly smaller – and consequently, the total energy required for the session also is. This is illustrated in Fig. 2.6 (bottom) where we see that for minimum expected delay routing most packets are delivered in a few hops and a few packets take a long time to be delivered. For fastest convergence rate routing, none of the packets took more than 8 hops to be delivered but the total number of hops required to deliver all the packets was larger.

2.3 A Saturated system approach

The approach in Section 2.2 ignores the effect of packet queuing at individual terminals. To incorporate this effect for heavily loaded networks, we consider that each user has an infinite-long queue. Packets arrive at random according to a Poisson process with rate $\rho_j \in (0, 1]$ to be delivered from terminal U_j to the destination U_{J+1}. We assume every user can transmit 1 packet per slot which implies that the service process is Poisson with rate $\mu_j = 1$. Our goal is to find conditions for the arrival rates ρ_j to yield stable queues and to design routing matrices $\mathbf{T}$ that maximize the sustainable ρ_j in some sense.

Besides its own packets, U_i receives packets from other nodes for an *aggregate* arrival

rate λ_i. A *necessary* condition for stable queues is $\lambda_i \leq \mu_i = 1$ in which case the departure process is also Poisson with rate λ_i (different from the service process whose rate is 1). If, as in Section 2.2, we let K_{ij} denote the probability that a packet moves from U_j to U_i between times n and $n+1$ we have that (see also Fig. 2.7)

$$\lambda_i = \rho_i + \sum_{j=1}^{J} K_{ij}\lambda_j, \tag{2.23}$$

where we used that a sum of Poisson processes is also a Poisson process. Notice that the sum in (2.23) includes the packets that fail to leave U_i in the term $K_{ii}\lambda_i$. Upon defining the vectors of (external) arrival rates $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_J]^T$ and aggregate arrival rates $\boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_J]^T$, we can express (2.23) in vector-matrix form as

$$\boldsymbol{\lambda} = \boldsymbol{\rho} + \mathbf{K}_0\boldsymbol{\lambda}, \tag{2.24}$$

with $\mathbf{K}_0$ denoting the $J \times J$ upper left corner of $\mathbf{K}$ as in (2.4). From a routing perspective, packets leave the network when they reach U_{J+1}, something that happens at a rate

$$\lambda_{J+1} = \sum_{j=1}^{J} K_{(J+1)j}\lambda_j = \mathbf{k}_1^T\boldsymbol{\lambda}. \tag{2.25}$$

Interestingly, we do not need the last column of $\mathbf{K}$ to describe this queuing model. For the remaining columns it is easy to see that the constraints are as in Section 2.2 and we thus look for routing matrices $\mathbf{T} \in \mathcal{T}$ for which we have $\mathbf{K} \in \mathcal{K}$ with $\mathcal{T}$ and $\mathcal{K}$ as in (2.5) and (2.6).

The first problem of interest is to find conditions under which pairs $(\mathbf{T}, \boldsymbol{\rho})$ of routing matrices and arrival rates are stable. Such a condition is given by the following theorem.

Theorem 4 *Suppose that arrivals adhere to Poisson processes with strictly positive rates given by $\boldsymbol{\rho}$ and the queue service rates are $\boldsymbol{\mu} = \mathbf{1}$. The queues at every user are stable if and only if $\mathbf{I} - \mathbf{K}_0$ is invertible and*

$$\boldsymbol{\lambda} = (\mathbf{I} - \mathbf{K}_0)^{-1}\boldsymbol{\rho} \preceq \mathbf{1}, \tag{2.26}$$

with $\preceq$ denoting componentwise inequality.

Proof: Each individual queue is $M/M/1$ with rates λ_j and $\mu_j = 1$. For $M/M/1$ queues, time reversibility implies independence of arrival and departure processes. We thus have that all arrival and departure processes are independent, the network of queues decouples and we have stability if and only if

$$\lambda_j \leq \mu_j = 1, \quad \forall j. \tag{2.27}$$

To show that (2.27) is equivalent to (2.26) it suffices to solve for $\boldsymbol{\lambda}$ in (2.24). This can be done when $\mathbf{I} - \mathbf{K}_0$ is invertible. To complete the proof we have to show that if $\mathbf{I} - \mathbf{K}_0$ is not invertible the queues are unstable.

To prove the latter note that since $\boldsymbol{\rho} \succ \mathbf{0}$ is componentwise strictly positive and all components K_{ij} of $\mathbf{K}_0$ are non-negative, then either $\mathbf{K}_0 = \mathbf{0}$ or $\mathbf{K}_0\boldsymbol{\rho} \neq \mathbf{0}$. In the first case $\mathbf{I} - \mathbf{K}_0 = \mathbf{I}$ is invertible; so, it must be that $\mathbf{K}_0\boldsymbol{\rho} \neq \mathbf{0}$. But if this is the case recursive application of (2.24) yields

$$\boldsymbol{\lambda} = \left(\sum_{n=0}^{\infty} \mathbf{K}_0^n \right) \boldsymbol{\rho}. \tag{2.28}$$

But non-invertibility of $\mathbf{I} - \mathbf{K}_0$ implies 1 is an eigenvalue of $\mathbf{K}_0$ and consequently $\rho(\mathbf{K}_0) \geq 1$; thus, the series in (2.28) diverges and we have that $\boldsymbol{\lambda}$ is arbitrarily large implying unstable queues. □

Theorem 4 provides a condition for having stable queues and in that sense it is the counterpart of (2.8). Given a routing matrix $\mathbf{T}$ and a vector of arrival rates $\boldsymbol{\rho}$, (2.28) can be used to check stability. For any candidate routing matrix $\mathbf{T}$, we can find the stability region $\mathcal{S}$ of arrival rate vectors leading to stable queues as

$$\mathcal{S} = \{\boldsymbol{\rho} \in \mathbb{R}^J : \boldsymbol{\rho} = (\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda}, \text{ with } \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}\}. \tag{2.29}$$

Perhaps more important than the stability region for a given routing matrix $\mathbf{T}$ is to design this matrix so that $\boldsymbol{\rho}$ is maximized in some sense. We pursue this problem in the next section.

2.3.1 Physical/medium-access/network layer interaction

Because the framework so far as well as the SRPs of the ensuing section rely on knowledge of $\mathbf{R}$, we delineate here how this matrix is determined depending on the access scheme (orthogonal or non-orthogonal) used at the physical layer.

If terminals transmit over orthogonal channels as when frequency (F-), time (T-), or code (C-) division multiple access (DMA) is utilized at the physical layer, $\mathbf{R}$ clearly depends on the power transmitted by individual users. Indeed, for the SNR in (2.1) and a given modulation and error control code pair one can readily obtain a certain packet success probability $R_{ij}(P_j)$ for the link $U_j \to U_i$. Depending on how fast fading varies with respect to packet lengths channels are classified as fast, slow, or block fading [31]. If fading is invariant over several packet transmissions, $R_{ij}(P_j)$ is given by the instantaneous packet success probabilities for the given fading state. If fading is fast, so that any packet experiences a sufficiently large number of independent channel realizations, the receiver can collect the available time diversity and $R_{ij}(P_j)$ can be approximately obtained from the error probability for additive white Gaussian noise channels. In a block fading model, the channel changes from packet to packet, and the transmitter is confronted with an unknown fading state. In this case $R_{ij}(P_j)$ can be computed from the average of the instantaneous error probabilities over all fading states. In all three cases, $R_{ij}(P_j)$ is expressible as a function of P_j.

For contention- or interference-limited networks as is respectively the case for random access and CDMA with pseudo-noise spreading sequences $\mathbf{R}$ and $\boldsymbol{\lambda}$ are coupled in the sense that $\mathbf{R}(\mathbf{p}, \boldsymbol{\lambda})$ is a function of the transmitted powers $\mathbf{p} := [P_1, \ldots, P_J]^T$ and the departure rates $\boldsymbol{\lambda}$. Indeed, reducing the transmission rate λ_j of U_j decreases the interference to other terminals consequently increasing the probability R_{ik} of U_i to successfully decode any U_k other than U_i. Since this coupling complicates matters substantially, a common approach is to assume for the purposes of accounting for interference that $\boldsymbol{\lambda} = \mathbf{1}$, which implies that the SINR is given by (2.2). Note that this eliminates $\boldsymbol{\lambda}$ as a variable determining $\mathbf{R}(\mathbf{p}, \mathbf{1})$ that now depends only on $\mathbf{p}$. Decoupling $\mathbf{R}$ from $\boldsymbol{\lambda}$ is an approximation tantamount to decoupling the networking from the medium access control (sub-) layer. The approximation

can be justified by noting that any rate $\boldsymbol{\rho}$ achievable in a network with reliability matrix $\mathbf{R}(\mathbf{p}, \boldsymbol{\lambda})$ is also achievable in a network with reliability $\mathbf{R}(\mathbf{p}, \mathbf{1})$ and in that sense the latter represents an upper bound on the stability region of the former.

2.3.2 Maximum arrival rate routing

Different optimization criteria can be devised to obtain routing algorithms maximizing the arrival rate vector. A first approach is to maximize a weighted sum of rates $\sum_{j=1}^{J} \rho_j = \mathbf{w}^T \boldsymbol{\rho}$ with $\mathbf{w} \succeq \mathbf{0}$. The sum-rate optimal matrix can be obtained as the solution of the optimization problem

$$\mathbf{K}^* = \arg \max_{\mathbf{K} \in \mathcal{K}, \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} \mathbf{w}^T \boldsymbol{\rho} = \max_{\mathbf{K} \in \mathcal{K}, \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} \mathbf{w}^T (\mathbf{I} - \mathbf{K}_0) \boldsymbol{\lambda}. \tag{2.30}$$

A concern with the formulation in (2.30) is that it tends to favor terminals close to the destination. An alternative approach is to maximize $\min_j \rho_j$, the rate of the least favored user. We refer to this as max-min optimal routing; the corresponding routing matrix can be obtained as the solution to

$$\mathbf{K}^* = \arg \max_{\mathbf{K} \in \mathcal{K}, \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} \min_j \rho_j = \max_{\mathbf{K} \in \mathcal{K}, \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} \min_j \left[(\mathbf{I} - \mathbf{K}_0) \boldsymbol{\lambda} \right]_j . \tag{2.31}$$

The optimization problems in (2.30) and (2.31) are bilinear in $\mathbf{K}_0$ and $\boldsymbol{\lambda}$, and as such, notoriously difficult to solve in general. Enticingly, we can capitalize on the structure of the problem to reduce them to simple linear programs. The main result allowing this reduction is stated in the following theorem.

Theorem 5 *Consider a maximization problem of the form*

$$v^* := \max_{\mathbf{K} \in \mathcal{K}, \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} g[(\mathbf{I} - \mathbf{K}_0) \boldsymbol{\lambda}], \tag{2.32}$$

where $g : \mathbb{R}^J \to \mathbb{R}$ is a function monotonically non-decreasing in each component, i.e., for vectors $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$ with $v_j^{(1)} \leq v_j^{(2)}$ and $v_i^{(1)} = v_i^{(2)}$ for $i \neq j$, we have that $g[\mathbf{v}^{(1)}] \leq g[\mathbf{v}^{(2)}]$. Then, there exists a matrix $\mathbf{K} \in \mathcal{K}$ such that

$$v^* = \max_{\mathbf{K} \in \mathcal{K}} g[(\mathbf{I} - \mathbf{K}_0) \mathbf{1}]. \tag{2.33}$$

Proof: Reasoning by contradiction, let $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ be a solution with $\lambda_j^{(1)} < 1$. Consider the alternative solution $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$ with $\lambda_j^{(2)} = 1$ and

$$T_{ij}^{(2)} = T_{ij}^{(1)} \lambda_j^{(1)}, \; i \neq J+1; \quad T_{(J+1)j}^{(2)} = 1 - \sum_{i=1}^{J} T_{ij}^{(2)}. \tag{2.34}$$

Since $0 \leq T_{ij}^{(1)} \leq 1$ and $0 \leq \lambda_j^{(1)} \leq 1$, we have that $0 \leq T_{ij}^{(1)} \leq 1$; furthermore, $T_{(J+1)j}^{(2)}$ are chosen so that $\sum_{i=1}^{J+1} T_{ij}^{(2)} = 1$ implying that $\left[\mathbf{T}^{(2)}\right]^T \mathbf{1} = \mathbf{1}$. Thus, the matrix $\mathbf{K}^{(2)}$ obtained from $\mathbf{T}^{(2)}$ by (3.1) is such that $\mathbf{K}^{(2)} \in \mathcal{K}$. Since by construction we also have $\mathbf{0} \preceq \boldsymbol{\lambda}^{(2)} \preceq \mathbf{1}$, we infer that $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$ is a feasible point of the optimization problem in (2.32). The proof relies on the following lemma.

Lemma 1 *Let $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$ with $\lambda_j^{(2)} = 1$ be obtained from $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ with $\lambda_j^{(1)} < 1$ from (2.36). If $\mathbf{K}_0^{(1)}$ and $\mathbf{K}_0^{(2)}$ denote the $J \times J$ upper left blocks of $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$, then*

$$g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}] \leq g[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]. \tag{2.35}$$

Proof: See Appendix B. □

If the inequality in (2.35) holds strictly, i.e., $g[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}] > g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]$, then Lemma 1 implies that $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ does *not* maximize the argument of (2.32) since at least one feasible point $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$ yields a larger argument. In this case the proof follows by contradiction.

If equality holds, i.e., $g[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}] = g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]$, we distinguish between $\boldsymbol{\lambda}^{(2)} = \mathbf{1}$ and $\boldsymbol{\lambda}^{(2)} \neq \mathbf{1}$. In the first case $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ may be an optimum solution, but if it is $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)}) = (\mathbf{1}, \mathbf{T}^{(2)})$ also is and the proof follows.

If $\boldsymbol{\lambda}^{(2)} \neq \mathbf{1}$, consider $(\boldsymbol{\lambda}^{(3)}, \mathbf{T}^{(3)})$ with $\lambda_k^{(3)} = 1$ and

$$T_{ik}^{(3)} = T_{ik}^{(2)} \lambda_k^{(2)}, \; i \neq J+1; \quad T_{(J+1)k}^{(3)} = 1 - \sum_{i=1}^{J} T_{ik}^{(3)}. \tag{2.36}$$

For this new solution, $\lambda_k^{(3)} = \lambda_j^{(3)} = 1$, and applying Lemma 1 we have

$$g[(\mathbf{I} - \mathbf{K}_0^{(3)})\boldsymbol{\lambda}^{(3)}] \geq g[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}] \geq g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]. \tag{2.37}$$

Repeating this argument, we build a succession of feasible points $(\boldsymbol{\lambda}^{(j)}, \mathbf{T}^{(j)})$ such that $g[(\mathbf{I} - \mathbf{K}_0^{(j)})\boldsymbol{\lambda}^{(j)}] \geq g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]$, for all j in which the number of components of $\boldsymbol{\lambda}^{(j)}$

equal to one is at least j. If at least one equality holds strictly, then $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ is *not* an optimum point. If no equality holds strictly, $g[(\mathbf{I} - \mathbf{K}_0^{(j)})\boldsymbol{\lambda}^{(j)}] = g[(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]$, for all j, and for some $j \leq J$, $\boldsymbol{\lambda}^{(j)} = \mathbf{1}$. In this latter case, if $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)})$ is an optimum point, then $(\boldsymbol{\lambda}^{(j)}, \mathbf{T}^{(j)}) = (\mathbf{1}, \mathbf{T}^{(j)})$ also is. □

Theorem 5 establishes that routing algorithms involving component-wise non-decreasing objective functions can be solved by setting $\boldsymbol{\lambda} = \mathbf{1}$ in the argument function to be optimized. Clearly, this is the case for max-min rate optimal and sum-rate optimal routing in which the functions are $g(\mathbf{v}) = \min_i(v_i)$ and $g(\mathbf{v}) = \mathbf{1}^T\mathbf{v}$, respectively. Furthermore, with $\boldsymbol{\lambda} = \mathbf{1}$, the bilinear arguments in (2.38) and (2.42) become linear functions of $\mathbf{K}_0$ implying the following corollary.

Corollary 1 *Max-min optimal routing and sum-rate optimal routing can be obtained as solutions of linear programs (LP) in* $\mathbf{K}$*:*

(i) For max-min optimal routing

$$\mathbf{K}^* = \arg\max_{\mathbf{K}\in\mathcal{K}} \min_i \left[(\mathbf{I} - \mathbf{K}_0)\mathbf{1}\right]_i . \tag{2.38}$$

(ii) For sum-rate optimal routing

$$\mathbf{K}^* = \arg\max_{\mathbf{K}\in\mathcal{K}} \mathbf{w}^T(\mathbf{I} - \mathbf{K}_0)\mathbf{1}. \tag{2.39}$$

Proof: The functions $g(\mathbf{v}) = \min_i(v_i)$ and $g(\mathbf{v}) = \mathbf{1}^T\mathbf{v}$ are component-wise non-decreasing in the sense considered in Theorem 5. This proves the equivalence of (2.38) with (2.31) and (2.42) with (2.30), respectively. That (2.42) is an LP follows after noting that the argument to be maximized is linear and recalling that the set $\mathcal{K}$ is a convex polyhedron. To prove that (2.38) is an LP introduce the auxiliary variable $w \geq \left[(\mathbf{I} - \mathbf{K}_0)\mathbf{1}\right]_i$, for all i and rewrite the maximization as

$$\begin{aligned} &\max \quad w \\ &\text{s.t.} \quad \mathbf{K} \in \mathcal{K}, \quad w\mathbf{1} \leq (\mathbf{I} - \mathbf{K}_0)\mathbf{1}. \end{aligned} \tag{2.40}$$

In (2.43), the argument and the constraints are linear entailing, by definition, an LP. □

Corollary 1 demonstrates that sum-rate and max-min optimal routing can be efficiently solved by convex optimization techniques, e.g., interior point methods [13]. Solving an LP incurs complexity $O(J^{3.5})$ [13, Ch. 11] in the number of nodes J and in that sense it is only moderately more complex than finding a traditional shortest path route whose complexity is $O(J^2)$ [8, Ch. 5].

Remark 2 For (2.33) to be true, every terminal must operate with aggregate departure rate $\lambda_j = 1$. The proof of Theorem 5 considers an operating point with $\lambda_j^{(1)} < 1$ and shows that increasing $\lambda_j^{(1)}$ to $\lambda_j^{(2)} = 1$ entails a larger set of stable arrival rates $\mathcal{S}^{(2)} > \mathcal{S}^{(1)}$, or, at the very least $\mathcal{S}^{(2)} = \mathcal{S}^{(1)}$. Comparing $\mathbf{T}^{(1)}$ with $\mathbf{T}^{(2)}$, we see that the strategy used by the proof is to send all the extra traffic $\lambda_j^{(2)} - \lambda_j^{(1)} = 1 - \lambda_j^{(1)}$ to the destination U_{J+1}. Indeed, writing $T_{(J+1)j}^{(\cdot)} = 1 - \sum_{i=1}^{J} T_{ij}^{(\cdot)}$ and using $T_{ij}^{(2)} = T_{ij}^{(1)} \lambda_j^{(1)}$ from (2.36), it follows after direct manipulation that

$$T_{(J+1)j}^{(2)} \lambda_j^{(2)} - T_{(J+1)j}^{(1)} \lambda_j^{(1)} = \lambda_j^{(2)} - \lambda_j^{(1)}. \tag{2.41}$$

The right hand side of (2.41) is precisely the traffic increase while the left hand side is the increase in traffic routed directly to the destination. This strategy will not increase the amount of traffic handled by other terminals $U_i \neq U_j$; while if $R_{(J+1)j} > 0$, it will increase the amount of traffic delivered by U_j. Furthermore, this argument also shows that $\mathcal{S}^{(2)} = \mathcal{S}^{(1)}$ if and only if $R_{(J+1)j} = 0$ – see also (2.71).

2.3.3 Commonly used rate optimality criteria

Requiring $f(\boldsymbol{\rho})$ to be monotonically non-decreasing in each component is a mild condition which ensures that an increase in the rate of one user does not decrease the value of the objective function to be maximized. Thus, the applicability of Theorem 5 is fairly broad, implying that we can propose different routing algorithms and expect them to yield tractable optimization problems. This include many "workhorse" optimality criteria that we describe next.

Optimal weighted sum with guaranteed rate. A variation of the weighted sum-rate criterion is to require a minimum acceptable rate $\rho_j^{\min}$ per terminal U_j. Upon defining the

vector $\boldsymbol{\rho}^{\min} := [\rho_1^{\min}, \ldots, \rho_J^{\min}]^T$, the optimal routing matrix is obtained as

$$\begin{aligned} \mathbf{K}^* = \arg\max \quad & \boldsymbol{\alpha}^T(\mathbf{I}-\mathbf{K}_0)\mathbf{1} \\ \text{s.t.} \quad & \mathbf{K} \in \mathcal{K},\ \boldsymbol{\rho}^{\min} \preceq (\mathbf{I}-\mathbf{K}_0)\mathbf{1}. \end{aligned} \tag{2.42}$$

Since $\mathcal{K}$ is a convex polyhedron, the constraint $\boldsymbol{\rho}^{\min} \preceq (\mathbf{I}-\mathbf{K}_0)\mathbf{1}$ is a set of linear inequalities and the objective $\boldsymbol{\alpha}^T(\mathbf{I}-\mathbf{K}_0)\mathbf{1}$ is linear, the optimization in (2.42) is a linear program (LP) in $\mathbf{K}$ and $\mathbf{T}$.

A solution $\mathbf{K}^*$ to (2.42) may not exist for some values of $\boldsymbol{\rho}^{\min}$ – in such cases interior point methods return an infeasibility certificate. When it exists, $\mathbf{K}^*$ ensures the minimum acceptable rate $\rho_j^{\min}$ to every user with the excess traffic distributed to the most favored users with large values of α_j and/or reliable connections to one of the APs.

Max-min optimal rate. As we established in Corollary 1 and repeat here for reference purposes the problem in (2.31) can also be rewritten as an LP. Indeed, note that $w \leq [(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}]_j$ for all j if and only if $w \leq \min_j [(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}]_j$; that is, the auxiliary variable w is smaller than the minimum rate if and only if it is smaller than all rates. Using this and the fact that $\min_j \rho_j$ is monotonically non-decreasing in each component (so that $\boldsymbol{\lambda}$ can be set to $\mathbf{1}$ without loss of generality) problem (2.31) can be recast as

$$\begin{aligned} \max \quad & w \\ \text{s.t.} \quad & \mathbf{K} \in \mathcal{K},\ \ w\mathbf{1} \leq (\mathbf{I}-\mathbf{K}_0)\mathbf{1} \end{aligned} \tag{2.43}$$

which is an LP in the auxiliary variable w and the problem variables $\mathbf{K}$ and $\mathbf{T}$. Max-min routing is fair in the sense that nodes in the network collaborate to optimize the rate of the worst user, pretty much along the spirit of max-min flow control [8, Section 6.5.2].

In the same way we added extra constraints to the optimal weighted sum-rate criterion [cf. (2.30) and (2.42)] we can add convex constraints to the rate vector $\boldsymbol{\rho}$ without altering the convexity of the problem. Another example of a convex constraint is a cooperation constraint whereby terminals require their own rate to be at least a certain percentage $\beta_j \in [0,1]$ of their total outgoing rate. The latter is given by the rate of packets successfully

transmitted to any terminal $\sum_{i=1,i\neq j}^{J+J_{\text{ap}}} K_{ij} = 1-K_{jj}$. The limit on the amount of cooperation can thus be enforced by adding the (convex) constraint $\beta_j \left(1 - \sum_{i=1}^{J} K_{ji}\right) \geq 1 - K_{jj}$.

Optimal rate with relays. In a relay network a group of terminals collaborate in relaying traffic on behalf of a designated active user. Let U_{j_0} denote this active user and terminals $\{U_j\}_{j=1,j\neq j_0}^{J}$ be the relays. The optimal relay network maximizing the rate of U_{j_0} can be found as

$$\begin{aligned} \mathbf{K}^* = \arg\max \quad & \left[(\mathbf{I}-\mathbf{K}_0)\mathbf{1}\right]_{j_0} \\ \text{s.t.} \quad & \mathbf{K}\in\mathcal{K},\ \left[(\mathbf{I}-\mathbf{K}_0)\mathbf{1}\right]_j = 0, \quad j\neq j_0. \end{aligned} \tag{2.44}$$

Indeed, notice that the constraints $\left[(\mathbf{I}-\mathbf{K}_0)\mathbf{1}\right]_j = 0$ for $j \neq j_0$ set the relay's traffic to zero, while the argument $\left[(\mathbf{I}-\mathbf{K}_0)\mathbf{1}\right]_{j_0}$ is the rate of the designated active user. The problem in (2.42) is, again, an LP in $\mathbf{K}$ and $\mathbf{T}$.

Optimal product of rates. Maximizing the product of rates constitutes a more fair alternative to the maximum sum-rate criterion in (2.30) since it prevents solutions in which some users receive a very small packet delivery rate. The function to be maximized in this case is $f(\boldsymbol{\rho}) = \prod_{j=1}^{J} \rho_j$ and the corresponding optimal routing matrix is obtained as

$$\mathbf{K}^* = \arg \max_{\mathbf{K}\in\mathcal{K},\mathbf{0}\preceq\boldsymbol{\lambda}\preceq\mathbf{1}} \prod_{j=1}^{J} \rho_j = \arg \max_{\mathbf{K}\in\mathcal{K},\mathbf{0}\preceq\boldsymbol{\lambda}\preceq\mathbf{1}} \sum_{j=1}^{J} \log\left[(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}\right]_j, \tag{2.45}$$

with the second equality following because the logarithm function is monotonically increasing. Notice that the argument in (2.45) is also monotonically non-decreasing in each component which allows one to set $\boldsymbol{\lambda} = \mathbf{1}$ without loosing optimality in the solution:

$$\mathbf{K}^* = \arg \max_{\mathbf{K}\in\mathcal{K}} \sum_{j=1}^{J} \log\left[(\mathbf{I}-\mathbf{K}_0)\mathbf{1}\right]_j. \tag{2.46}$$

Since the logarithm is a concave function, (2.46) is a convex optimization problem in $\mathbf{K}$ and $\mathbf{T}$. Thus, globally convergent interior point methods can be readily applied here too.

2.3.4 An overall constraint in the total traffic

Imposing individual traffic constraints, the requirement $\mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}$ does not impose an overall traffic constraint, something that is sometimes reasonable and sometimes not. In

certain cases we may want to limit the total traffic in the network, e.g., to leave room for critical traffic, or, to ensure a fixed power consumption per time unit. In any event, the total traffic constraint can be written, without loss of generality, as $\boldsymbol{\lambda}^T\mathbf{1} = 1$ – any constant other than 1 could be used. In this context, we can consider different optimization criteria as in Section 2.3.2 yielding routing algorithms of the form

$$\begin{aligned}\mathbf{K}^* = \arg\max_{\mathbf{K},\boldsymbol{\lambda}} \;& g\left[(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}\right] \\ \text{s.t.} \;& \mathbf{K}\in\mathcal{K},\; \mathbf{0}\preceq\boldsymbol{\lambda},\; \boldsymbol{\lambda}^T\mathbf{1}=1.\end{aligned} \tag{2.47}$$

The added constraint $\boldsymbol{\lambda}^T\mathbf{1} = 1$ prevents application of Theorem 5 and, in general, problems of the form (2.47) will be difficult to solve. However, for the specific case of max-min optimal routing with an overall traffic constraint, i.e., $g(\mathbf{v}) = \min_i(v_i)$ in (2.47), we can establish a quite surprising connection with shortest path routing.

To study this connection note that since the constraints in $\boldsymbol{\lambda}$ and $\mathbf{K}_0$ are decoupled we can solve the optimization in two separate steps

$$\mathbf{K}^* = \arg\max_{\mathbf{K}\in\mathcal{K}} \left\{ \begin{array}{ll} \max_{\boldsymbol{\lambda}} & \min_i\left[(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}\right]_i \\ \text{s.t.} & \mathbf{0}\preceq\boldsymbol{\lambda},\; \boldsymbol{\lambda}^T\mathbf{1}=1 \end{array} \right\}. \tag{2.48}$$

If $\mathbf{K}_0$ is fixed, then the innermost optimization is a simple linear max-min problem widely studied in a variety of contexts, e.g., game theory [63, chap.2]. The important point here is that the solution to this problem is well known, and in some cases computable in closed-form. This allows us to obtain the following theorem.

Theorem 6 *For consistent routing matrices, max-min optimal routing with a global traffic constraint as defined by (2.48) is equivalent to*

$$\mathbf{K}^* = \arg\min_{\mathbf{K}\in\mathcal{K}} \mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-1}\mathbf{1}. \tag{2.49}$$

Proof: Consider the innermost maximization in (2.48) and let u^* denote the optimum value

$$\begin{aligned}u^* = \max_{\boldsymbol{\lambda}} \;& \min_i\left[(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\lambda}\right]_i \\ \text{s.t.} \;& \mathbf{0}\preceq\boldsymbol{\lambda},\; \boldsymbol{\lambda}^T\mathbf{1}=1,\end{aligned} \tag{2.50}$$

that is achieved by the vector $\boldsymbol{\lambda}^*$. The maximum u^* can be found in closed-form, and for that matter consider a vector $\boldsymbol{\lambda}^\dagger$ such that

$$(\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda}^\dagger = u^\dagger \mathbf{1}. \tag{2.51}$$

In order for $\boldsymbol{\lambda}^\dagger$ to be a feasible point of the problem in (2.50), it is necessary to have $\mathbf{1}^T\boldsymbol{\lambda}^\dagger = 1$ from where we obtain

$$u^\dagger = \frac{1}{\mathbf{1}^T(\mathbf{I} - \mathbf{K}_0)^{-1}\mathbf{1}}. \tag{2.52}$$

The corresponding $\boldsymbol{\lambda}^\dagger$ can then be found as

$$\boldsymbol{\lambda}^\dagger = \frac{(\mathbf{I} - \mathbf{K}_0)^{-1}\mathbf{1}}{\mathbf{1}^T(\mathbf{I} - \mathbf{K}_0)^{-1}\mathbf{1}}. \tag{2.53}$$

The key observation now is that for consistent routing matrices all the components of $(\mathbf{I} - \mathbf{K}_0)^{-1}$ are positive. Indeed, if $\mathbf{K}_0$ is consistent then $(\mathbf{I} - \mathbf{K}_0)^{-1} = \sum_{n=1}^{\infty} \mathbf{K}_0^n$. But since $K_{ij} \geq 0$ for all i, j, we have that all the components of $(\mathbf{I} - \mathbf{K}_0)^{-1}$ also are. We thus infer that $\boldsymbol{\lambda}^\dagger \succeq \mathbf{0}$, a fact that combined with $\mathbf{1}^T\boldsymbol{\lambda}^\dagger = 1$ [which was enforced in deriving (2.52)] implies that $\boldsymbol{\lambda}^\dagger$ is a feasible point of the problem in (2.50). For this feasible point, we have that

$$\min_i \left[(\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda}^\dagger\right]_i = u^\dagger, \tag{2.54}$$

since $[(\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda}^\dagger]_i = u^\dagger$, for all i.

Consider now the dual optimization problem of (2.50) which can be obtained by writing the Lagrangian function or simply recalling Von-Neumann's min-max theorem. Either way, we obtain

$$\begin{aligned} v^* = \min_{\boldsymbol{\mu}} \quad & \max_i \left[(\mathbf{I} - \mathbf{K}_0)^T\boldsymbol{\mu}\right]_i \\ \text{s.t.} \quad & \mathbf{0} \preceq \boldsymbol{\mu},\ \boldsymbol{\mu}^T\mathbf{1} = 1, \end{aligned} \tag{2.55}$$

with the corresponding maximizing argument denoted by $\boldsymbol{\mu}^*$. But note that (2.50) and (2.55) are quite similar, the only differences being the inversion of the maximum and minimum operators and the transposition of $(\mathbf{I} - \mathbf{K}_0)$. We thus find that the vector $\boldsymbol{\mu}^\dagger$ given

by

$$\boldsymbol{\mu}^\dagger = \frac{(\mathbf{I}-\mathbf{K}_0)^{-T}\mathbf{1}}{\mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-T}\mathbf{1}}, \tag{2.56}$$

is a feasible point of the problem in (2.55) such that

$$\max_i \left[(\mathbf{I}-\mathbf{K}_0)\boldsymbol{\mu}^\dagger\right]_i = \frac{1}{\mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-T}\mathbf{1}} =: v^\dagger. \tag{2.57}$$

Note that $v^\dagger = u^\dagger$ and that both values are achieved by respective feasible points $\boldsymbol{\lambda}^\dagger$ and $\boldsymbol{\mu}^\dagger$ of the primal problem in (2.50) and the dual problem in (2.55). Moreover, note that the problem in (2.50) is convex; it thus follows from the duality principle that $v^* = u^*$. On the other hand, the fact that u^* maximizes (2.50) and v^* minimizes (2.55) respectively imply that $u^\dagger \leq u^*$ and $v^* \leq v^\dagger$. Taken together, these yield

$$u^\dagger \leq u^* = v^* \leq v^\dagger \tag{2.58}$$

where the first inequality follows because u^* maximizes (2.50), the second equality from the duality principle for convex optimization problems, and the third inequality is true because v^* solves (2.55).

Upon comparing (2.52) with (2.57) we can see that $u^\dagger = v^\dagger$; hence, it must hold that $v^* = v^\dagger = u^\dagger = u^*$ [cf. (2.58)] and we can now rewrite (2.48) as

$$\mathbf{K}^* = \arg\max_{\mathbf{K}\in\mathcal{K}} u^* = \arg\max_{\mathbf{K}\in\mathcal{K}} \frac{1}{\mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-1}\mathbf{1}} \tag{2.59}$$

But since u^* is always positive – all components of $(\mathbf{I}-\mathbf{K}_0)^{-1}$ are –, maximizing u^* is equivalent to minimizing $1/u^*$ and (2.49) follows. □

Even though Theorem 6 transforms the problem in (2.48) into a conceptually simpler form, it is not yet clear how (2.49) might be solved. However, recalling (2.22) we see that quite surprisingly, max-min rate routing with a global traffic constraint as per (2.48) is equivalent to minimum delay routing as defined in (2.19). Since the solution to the latter, as we have already seen, is given by the shortest path in a fully connected graph with arc weights $1/R_{ij}$, so is the solution to (2.48); a fact that we summarize in the following corollary.

Corollary 2 *The matrix* $\mathbf{K}^\dagger \in \mathcal{K}$ *satisfying Bellman's principle of optimality in (2.20) solves the max-min routing problem with a global traffic constraint defined by (2.48).*

Proof: If $\mathbf{K}^\dagger \in \mathcal{K}$ satisfies (2.20), it solves (2.22) [cf. Theorem 3]. But (2.22) is identical to (2.49) which we know solves (2.48) [cf. Theorem 6]. Thus, if $\mathbf{K}^\dagger \in \mathcal{K}$ satisfies (2.20), it solves (2.48). □

Corollary 2 implies that in order to find the matrix optimizing (2.48) it suffices to run Algorithm 1. On the other hand, the proof of Theorem 6 provides interesting insights on the optimal solution that we discuss in the following remarks

Remark 3 The proof establishes that for any $\mathbf{K}_0$, the optimal $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\dagger$ is given by (2.53). The corresponding rate offered to each user is subsequently given by $\rho_j = 1/\mathbf{1}^T(\mathbf{I}-\mathbf{K}_0)^{-1}\mathbf{1}$ as stated in (2.52) showing that every user gets the same rate. In particular, this is true for $\mathbf{K}_0 = \mathbf{K}_0^*$ implying that the vector of optimal offered rates is

$$\boldsymbol{\rho}^* = \left[\mathbf{1}^T(\mathbf{I}-\mathbf{K}_0^*)^{-1}\mathbf{1}\right]^{-1}\mathbf{1}. \tag{2.60}$$

Eq. (5.99) reveals that max-min routing is fair in the sense that it evenly divides the traffic resources available.

Remark 4 Strong duality applied to the innermost optimization over $\boldsymbol{\lambda}$ in (2.48) proves the equivalence of (2.50) with (2.55). We can thus rewrite (2.48) as

$$\mathbf{K}^* = \arg\max_{\mathbf{K}\in\mathcal{K}} \left\{ \begin{array}{ll} \min_{\boldsymbol{\mu}} & \max_i \left[(\mathbf{I}-\mathbf{K}_0^T)\boldsymbol{\mu}\right]_i \\ \text{s.t.} & \mathbf{0} \preceq \boldsymbol{\mu},\ \boldsymbol{\mu}^T\mathbf{1} = 1 \end{array} \right\} \tag{2.61}$$

where we used that $(\mathbf{I}-\mathbf{K}_0)^T = \mathbf{I}-\mathbf{K}_0^T$. The formulation in (2.61) corresponds to min-max optimal routes for a multi-hop cooperative downlink subject to a constraint in the total traffic delivered by $D \equiv U_{J+1}$. The interpretation is that of a group of terminals competing to receive information from $D \equiv U_{J+1}$ that can transmit at a rate of 1 packet per packet slot [cf. $\boldsymbol{\mu}^T\mathbf{1} \leq 1$]. The access point (D) is interested in a fair formulation that minimizes the rate of the greediest user node while still using its own resources to a full extent [cf. $\boldsymbol{\mu}^T\mathbf{1} = 1$]. This problem turns out equivalent to max-min optimal routing for a multi-hop

cooperative uplink [cf. (2.48) and (2.61)]. In particular, we deduce that every node is served with the same rate given by (5.99).

2.4 Infrastructure with multiple APs

For the most part of this chapter we have worked with a single AP. This is more a matter of notational simplicity than of real necessity, since extensions to a network with multiple APs, i.e., with $J_{\text{ap}} > 1$, are straightforward.

Consider a network with J_{ap} APs and utilize $\mathbf{K}$ to capture the evolution of packets through the network as in Section 2.2. Repeating the steps leading to (2.4) we can see that none of them relies in the existence of a single AP. The only modification to the argument is that instead of requiring $T_{i(J+1)} = 0$, $\forall i \in [1, J]$ we require this to be true for all of the APs, consequently, we have that $T_{ij} = 0$, $\forall i \in [1, J + J_{\text{ap}}], i \neq j$, with the latter equation valid $\forall j \in [J + 1, J + J_{\text{ap}}]$. Thus, the matrix $\mathbf{K}$ can be partitioned as

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_0 & \mathbf{0} \\ \mathbf{K}_1^T & \mathbf{I} \end{pmatrix}_{(J+J_{\text{ap}})\times(J+J_{\text{ap}})}, \tag{2.62}$$

where, as before $\mathbf{K}_0$ denotes the $J \times J$ upper left submatrix of $\mathbf{K}$, $\mathbf{K}_1$ the $J_{\text{ap}} \times J$ lower left submatrix of $\mathbf{K}$, and $\mathbf{I}$ the $J_{\text{ap}} \times J_{\text{ap}}$ identity matrix. We would now have J_{ap} absorbing states and Theorem 1 has to be modified accordingly. The formulation of optimal routing algorithms depends on $\mathbf{K}_0$ only [cf. (2.11), (2.14), and (2.22)] and applies verbatim when $J_{\text{ap}} > 1$.

For the saturated system approach of Section 2.3 the extension is even simpler. Note that in deriving (2.24) we did not make use of the existence of a single AP. Thus arrival rates and aggregate arrival rates are related as in (2.24) that we repeat here for convenience,

$$\boldsymbol{\lambda} = \boldsymbol{\rho} + \mathbf{K}_0 \boldsymbol{\lambda} \tag{2.63}$$

The change is in the rate at which packets are delivered to the AP that in this case is given by

$$\boldsymbol{\lambda}_{\text{ap}} = \mathbf{K}_1 \boldsymbol{\lambda} \tag{2.64}$$

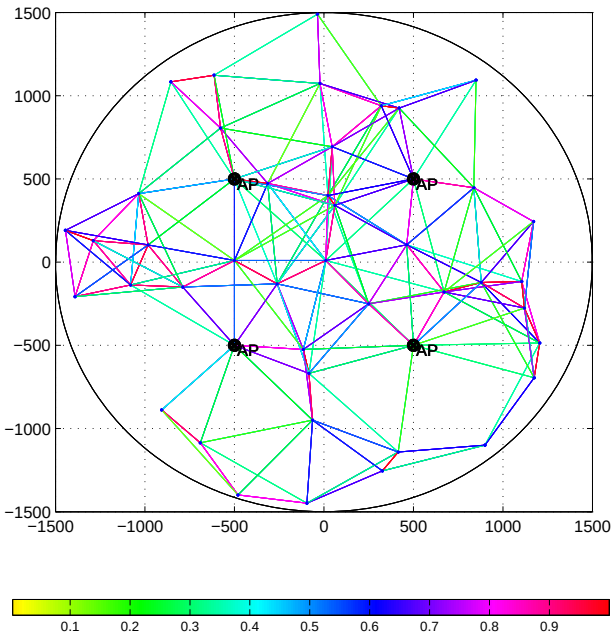

Figure 2.8: Schematic representation of the reliability matrix $\mathbf{R}$ for the network used in the simulations in Section 2.4.1. (R_{ij} is generated according to the empirical distribution in [22]; only arcs with $R_{ij} > 0.3$ are shown.)

with $\mathbf{K}_1$ as in (2.62). Note that the stability condition, as well as all routing problems formulated depend on $\mathbf{K}_0$ only and thus also apply verbatim to a network with $J_{\text{ap}} > 1$. The existence of many APs enters the problem formulations in (2.42)-(2.46) and (2.47) through the definition of the set $\mathcal{K}$.

2.4.1 Simulations and numerical examples

To illustrate the differences between the various routing protocols we consider the network with $J = 40$ nodes and $K = 4$ APs schematically represented in Fig. 2.8. The results of sum-rate optimal routing in (2.30) and max-min optimal routing as per (2.31) are shown in Figs. 2.9 and 3.1, respectively.

Sum-rate optimal routing yields a matrix $\mathbf{K}$ in which the nodes with reliable links to the destination are allocated most of the rate. Actually, a possible solution maximizing the sum-rate is for all the U_j's that are decoded by some AP with non-zero probability to

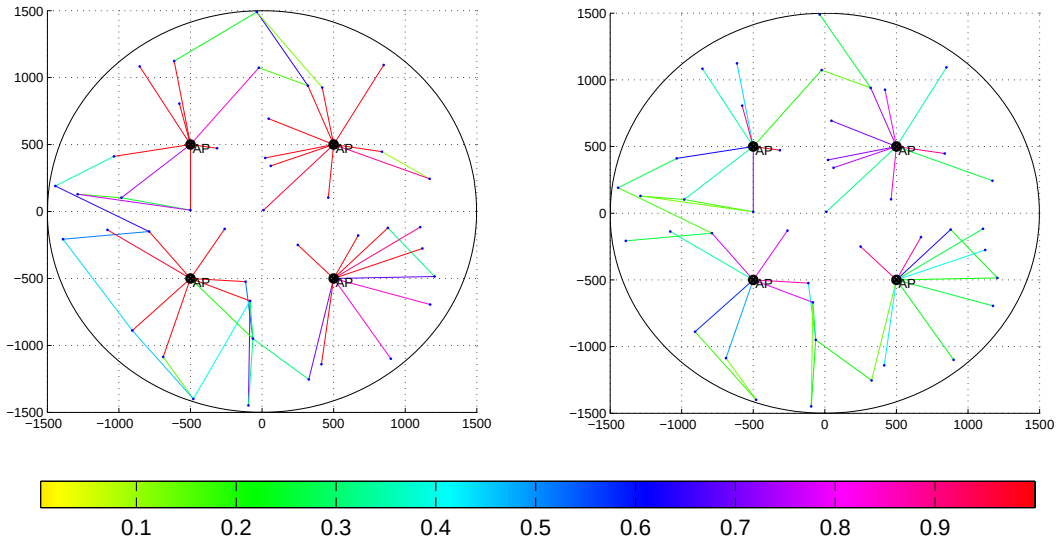

Figure 2.9: Sum-rate optimal routes with minimum acceptable rate as given by (2.42) for the network in Fig.2.8; matrices $\mathbf{T}$ (left) and $\mathbf{K}$ (right) are shown for $\mathbf{w} = \mathbf{1}$ and $\boldsymbol{\rho}^{\min} = 0.1\mathbf{1}$. Nodes with good connections to the destination get most of the total rate available.

send their traffic to that AP without forwarding any traffic belonging to other users. To this end, we add the constraint $\boldsymbol{\rho} \geq \boldsymbol{\rho}^{\min}$, which ensures that every user has a guaranteed rate $\rho_j = \rho_j^{\min}$ with the excess traffic assigned to the most favored users. The result of this approach is shown in Fig. 2.9 with $\boldsymbol{\rho}^{\min} = (0.1)\mathbf{1}$.

The optimal routes for the max-min criterion are depicted in Fig. 3.1. We see that most users divide their traffic between many different neighbors to avoid the formation of bottlenecks. The fairness of this approach is illustrated in Fig. 2.11 where instances of the arrival processes of the best and worst users are shown. For the network considered, the offered rates were 0.12 and 0.17, respectively. We see that the simulated arrival processes are accurately modeled by (2.23). We also plot the sum-rate for this case which is to be compared with 7.23 achieved by sum-rate optimal routing.

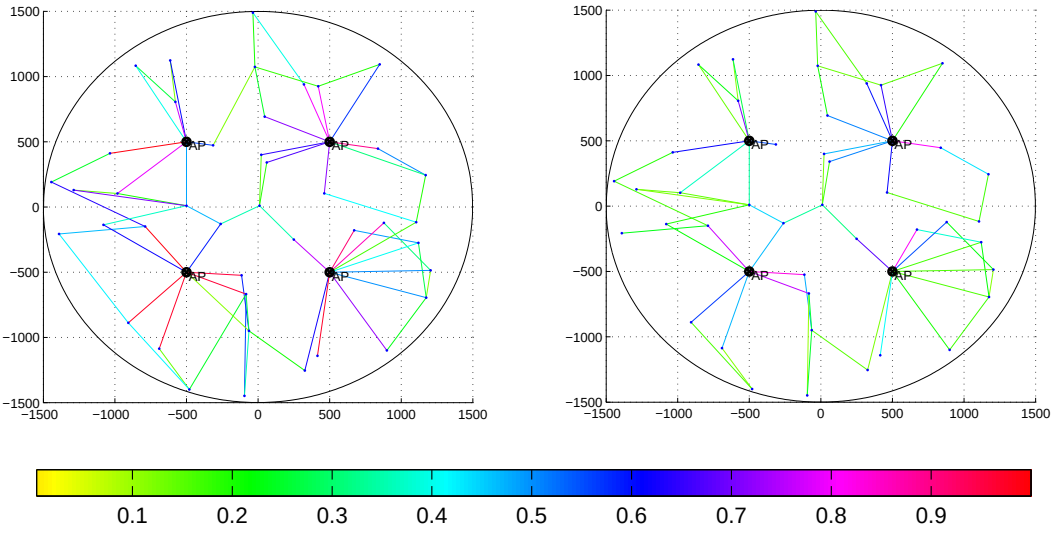

Figure 2.10: Max-Min routes obtained as the solution of (2.31) for the network in Fig.2.8; matrices $\mathbf{T}$ (left) and $\mathbf{K}$ (right) are shown. Compromised nodes divide their traffic among many different neighbors to avoid the formation of bottlenecks.

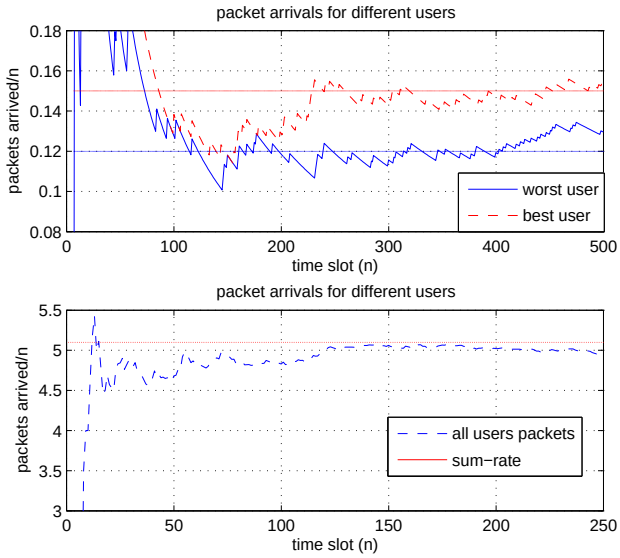

Figure 2.11: Instances of the arrival rate processes for the max-min optimal routes in Fig. 3.1. The fairness of the protocol is manifested in the not so different rates offered to the best and worst nodes.

2.5 Summary

We introduced a general framework for stochastic routing in wireless multi-hop networks. Deviating from the traditional graph models, we considered a general framework based on the packet delivery probability matrix and showed that different routing algorithms can be either described by the evolution of a properly defined Markov chain – per session model of routing –, or by a network of backlogged queues – saturated system. These connections permit characterization of properly defined deliverability and stability conditions in terms of the spectral radius of a stochastic routing matrix.

For the per-session model of routing we introduced stochastic routing algorithms that maximize the convergence rate of the Markov chain, entailing a maximization of the packet delivery probability for a fixed, sufficiently large delay n. This routing approach is meaningful in the context of delay sensitive traffic involved in, e.g, voice and/or video conferencing. We further found an expression for the average packet delay measured by the number of hops and identified the corresponding optimal routing scheme that minimizes it. Interestingly, we proved that the optimum routing matrix in this case can be obtained as the shortest path route in a fully connected graph with the arc between users having a weight inversely proportional to the corresponding delivery ratio.

For the saturated system we defined different routing algorithms corresponding to different maximization criteria of the arrival rate vector. These approaches include maximization of the rate of the least favored user (max-min), maximization of the sum of rates (max-sum) and of the product of rates (max-prod). We showed that all these problems can be efficiently solved using convex optimization techniques. Rather unexpectedly, we also established equivalence between max-min routing with a global traffic constraint and minimum average delay routing.

A problem the reader may have foreseen is that in order to obtain optimal routes the matrix $\mathbf{R}$ has to be collected at a central location and the optimal routing matrix $\mathbf{T}$ percolated through the network. An interesting problem is to develop *distributed* algorithms that find optimal routing probabilities without the burden of collecting $\mathbf{R}$ at a central node and then percolating the resulting routing probabilities through network nodes. We say an

algorithm is distributed if: i) terminal U_j has access only to the j-th row and column of $\mathbf{R}$; and ii) U_j interchanges variables only with those "one-hop neighbors" having positive probability of decoding its packets. The distributed algorithms can be built by recasting the optimization problems and applying dual decomposition techniques as in, e.g., [18, 57, 95]. Such algorithms are pursued in the next chapter.

Appendix

2.5.1 Proof of Theorem 3

Given $\bar{\delta}_i$ for $i \neq j$ we solve (2.16) for $\bar{\delta}_j$ to obtain,

$$\bar{\delta}_j = \frac{1 + \sum_{i=1, i\neq j}^{J+1} K_{ij}\bar{\delta}_i}{1 - K_{jj}}. \tag{2.65}$$

Since $\sum_{i=1}^{J+1} K_{ij} = 1$ we have that $1 - K_{jj} = \sum_{i=1, i\neq j}^{J+1} K_{ij}$; if we also replace $K_{ij} = T_{ij}R_{ij}$ valid for $i \neq j$ we obtain

$$\bar{\delta}_j = \frac{1 + \sum_{i=1, i\neq j}^{J+1} T_{ij}R_{ij}\bar{\delta}_i}{\sum_{i=1, i\neq j}^{J+1} T_{ij}R_{ij}}. \tag{2.66}$$

Now, replace the 1 in the numerator by $\sum_{i=1, i\neq j}^{J+1} T_{ij} = 1$ and rearrange terms to arrive at

$$\bar{\delta}_j = \frac{\sum_{i=1, i\neq j}^{J+1} \left(1/R_{ij} + \bar{\delta}_i\right) T_{ij}R_{ij}}{\sum_{i=1, i\neq j}^{J+1} T_{ij}R_{ij}}. \tag{2.67}$$

It also follows by definition that $(1/R_{ij}) + \bar{\delta}_i \geq \min_i \left(1/R_{ij} + \bar{\delta}_i\right)$ which allows us to bound $\bar{\delta}_j$ in (2.69) by

$$\begin{aligned} \bar{\delta}_j &\geq \min_i \left(\frac{1}{R_{ij}} + \bar{\delta}_i\right) \frac{\sum_{i=1, i\neq j}^{J+1} T_{ij}R_{ij}}{\sum_{i=1, i\neq j}^{J+1} T_{ij}R_{ij}} \\ &= \min_i \left(\frac{1}{R_{ij}} + \bar{\delta}_i\right). \end{aligned} \tag{2.68}$$

The matrix satisfying (2.20) for all j achieves the lower bound in (2.68) and thus minimizes $\bar{\delta}_j$ for all j. This proves that if a matrix satisfies (2.20) it minimizes $\bar{\delta}_j$ for all j. That such a matrix exists follows from the construction in Algorithm 1 that yields a matrix satisfying (2.20) as long as ensuring deliverability is possible.

For an arbitrary initial distribution we have that

$$\bar{\delta}[\mathbf{p}(0)] = \sum_{j=1}^{J} \Pr\{e_j(0)\}\bar{\delta}_j = \mathbf{p}^T(0)\bar{\boldsymbol{\delta}}. \tag{2.69}$$

But since all components $p_j(0)$ of $\mathbf{p}(0)$ are non-negative, $\bar{\delta}[\mathbf{p}(0)]$ is minimized if all components of $\bar{\boldsymbol{\delta}}$ are minimum. The latter is true if (2.20) is valid for all j [cf, (2.68)], completing the proof. □

2.5.2 Proof of Lemma 1

To see why (2.35) is true note that $(\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$ is such that the product $K_{ij}^{(2)}\lambda_j^{(2)} = K_{ij}^{(1)}\lambda_j^{(1)}$ for $i \neq J+1$ and $i \neq j$. Indeed,

$$K_{ij}^{(2)}\lambda_j^{(2)} = K_{ij}^{(2)} = R_{ij}T_{ij}^{(2)} = K_{ij}^{(1)}\lambda_j^{(1)},\ i \neq J+1, i \neq j, \tag{2.70}$$

where in the first equality we used that $\lambda_j^{(2)} = 1$.

For $i = J+1$ we have that the respective products are related by

$$\begin{aligned} K_{(J+1)j}^{(2)}\lambda_j^{(2)} &= R_{(J+1)j}\left[1 - \sum_{i=1}^{J} T_{ij}^{(2)}\right]\lambda_j^{(2)} \\ &= R_{(J+1)j}\left[\lambda_j^{(2)} - \sum_{i=1}^{J} T_{ij}^{(1)}\lambda_j^{(1)}\right] \\ &\geq R_{(J+1)j}\left[\lambda_j^{(1)} - \sum_{i=1}^{J} T_{ij}^{(1)}\lambda_j^{(1)}\right] \\ &= K_{(J+1)j}^{(1)}\lambda_j^{(1)} \end{aligned} \tag{2.71}$$

where: i) in the first equality we substituted $K_{(J+1)j}^{(2)} = R_{(J+1)j}T_{(J+1)j}^{(2)}$ and $T_{(J+1)j}^{(2)}$ for its expression in (2.36); ii) in the second one, we used $T_{ij}^{(2)}\lambda_j^{(2)} = T_{ij}^{(1)}\lambda_j^{(1)}$ that follows from (2.36) and $\lambda_j^{(2)} = 1$; iii) the inequality follows from $\lambda_j^{(1)} < \lambda_j^{(2)} = 1$; and iv) the last equality reverses the steps in i) by using $T_{(J+1)j}^{(1)} = 1 - \sum_{i=1}^{J} T_{ij}^{(1)}$ and $K_{(J+1)j}^{(2)} = R_{(J+1)j}T_{(J+1)j}^{(2)}$.

Consider the i^{th} component of the bilinear product in (2.70) $[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_i$. Since for $k \neq j$, $\lambda_k^{(2)} = \lambda_k^{(1)}$ and $K_{ik}^{(2)} = K_{ik}^{(1)}$, we have that

$$\begin{aligned} [(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_i &= \lambda_i^{(2)} - \sum_{k=1,k\neq j}^{J} K_{ik}^{(2)}\lambda_k^{(2)} - K_{ij}^{(2)}\lambda_j^{(2)} \\ &= \lambda_i^{(2)} - \sum_{k=1,k\neq j}^{J} K_{ik}^{(1)}\lambda_k^{(1)} - K_{ij}^{(2)}\lambda_j^{(2)}. \end{aligned} \tag{2.72}$$

For $i \neq j$ we can use (2.70) to show that $[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_i = [(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(1)}]_i$, i.e., that the i^{th} component of the bilinear product is invariant in the transformation $(\boldsymbol{\lambda}^{(1)}, \mathbf{T}^{(1)}) \rightarrow (\boldsymbol{\lambda}^{(2)}, \mathbf{T}^{(2)})$. Indeed, according to (2.70), when $i \neq j$, $K_{ij}^{(2)}\lambda_j^{(2)} = K_{ij}^{(1)}\lambda_j^{(1)}$ yielding [cf.

(2.70), (2.72), and $\lambda_i^{(2)} = \lambda_i^{(1)}$ for $i \neq j$]

$$[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_i = \lambda_i^{(1)} - \sum_{k=1,k\neq j}^{J} K_{ik}^{(1)}\lambda_k^{(1)} - K_{ij}^{(1)}\lambda_j^{(1)}$$
$$= [(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]_i, \tag{2.73}$$

thus proving the former claim.

For $i = j$, we can reorder terms in (2.72) to obtain

$$[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_j = -\sum_{k=1,k\neq j}^{J} K_{jk}^{(1)}\lambda_k^{(1)} + \lambda_j^{(2)}\left(1 - K_{jj}^{(2)}\right)$$
$$= -\sum_{k=1,k\neq j}^{J} K_{jk}^{(1)}\lambda_k^{(1)} + \sum_{k=1,k\neq j}^{J+1} K_{kj}^{(2)}\lambda_j^{(2)}, \tag{2.74}$$

where in the second equality we used that $\sum_{k=1}^{J+1} K_{kj}^{(2)} = 1$. But note that in the second summation in (2.72) we have that: i) for $k \neq J+1$, $K_{kj}^{(2)}\lambda_j^{(2)} = K_{kj}^{(1)}\lambda_j^{(1)}$ as stated in (2.70); and ii) for $k = J+1$ $K_{(J+1)j}^{(2)}\lambda_j^{(2)} \geq K_{(J+1)j}^{(1)}\lambda_j^{(1)}$ as stated in (2.71). Using these in (2.74) we obtain

$$[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_j \geq -\sum_{k=1,k\neq j}^{J} K_{jk}^{(1)}\lambda_k^{(1)} + \sum_{k=1,k\neq j}^{J+1} K_{kj}^{(1)}\lambda_j^{(1)}. \tag{2.75}$$

We now use again that $\sum_{k=1,k\neq j}^{J+1} K_{kj}^{(1)} = 1 - K_{jj}^{(1)}$ to reduce (2.75) to

$$[(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]_j \geq \lambda_j^{(1)} - \sum_{k=1,k\neq j}^{J} K_{jk}^{(1)}\lambda_k^{(1)} - K_{jj}^{(1)}\lambda_j^{(1)}$$
$$= [(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]_j. \tag{2.76}$$

The expressions in (2.73) and (2.76) imply that the vectors $\mathbf{v}^{(1)} := [(\mathbf{I} - \mathbf{K}_0^{(1)})\boldsymbol{\lambda}^{(1)}]$ and $\mathbf{v}^{(2)} := [(\mathbf{I} - \mathbf{K}_0^{(2)})\boldsymbol{\lambda}^{(2)}]$ are such that $v_j^{(1)} \leq v_j^{(2)}$ and $v_i^{(1)} = v_i^{(2)}$ for $i \neq j$. We thus use the componentwise monotonicity of $g(\cdot)$ to obtain (2.35).

Chapter 3

Self-organizing stochastic routing

Unlike the crude connectivity graph, the reliability matrix $\mathbf{R}$ – whose (i,j)-th entry R_{ij} represents the probability that a packet transmitted from the j-th user U_j is correctly received by the i-th user U_i – takes link reliability into account. It thus provides for a more accurate model of the wireless channel; however, the usefulness of a model based on $\mathbf{R}$ hinges on the relative communication efficiency and algorithmic complexity of finding optimal routes. Enticingly, among the routing protocols introduced in Chapter 2, the rate maximizing approaches of Section 2.3.2 lead to optimal routing algorithms in the form of convex optimization problems.

Recapitulating the results in Section 2.3 consider a wireless network with $J+1$ user nodes $\{U_j\}_{j=1}^{J+1}$ in which the first J users $\{U_j\}_{j=1}^{J}$ participate in routing packets to a destination $D \equiv U_{J+1}$. The physical and medium access layers are such that (s.t.) if a packet is transmitted by U_j it is correctly *received* by U_i with probability R_{ij} that we arrange in the matrix $\mathbf{R}$. Packets are stochastically routed according to probabilities T_{ij} collected in a matrix $\mathbf{T}$. When a user terminal U_j decides to transmit a packet it selects a random terminal as the intended destination with U_i chosen with probability T_{ij}. If the transmission is successfully received – something that happens with probability R_{ij} – the packet moves to U_i's queue; otherwise it is kept by U_j that attempts transmission, possibly to a different node, at a later time. To capture the evolution of packets through the network we define a matrix $\mathbf{K}$ whose elements K_{ij} represent the probability that a packet moves from U_j's

queue to U_i's queue. For $i \neq j$ the packet moves from U_j to U_i if and only if it is routed through U_i *and* is correctly decoded; since these two events are independent we have,

$$K_{ij} = T_{ij}R_{ij} \quad \text{for } i \neq j, \qquad \mathbf{K}^T\mathbf{1} = \mathbf{1}, \qquad \mathbf{T}^T\mathbf{1} = \mathbf{1} \tag{3.1}$$

where the last two constraints come from the fact that $\mathbf{K}$ and $\mathbf{T}$ are stochastic matrices.

To complete the formulation of the routing problem let $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_J]^T$ denote the vector of packet arrival rates and $\boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_J]^T$ the rate of departures that we constrain by $\mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}$. Defining $\mathbf{K}_0$ as the $J \times J$ upper left submatrix of $\mathbf{K}$ it is not difficult to see that we can relate $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ by [cf. (2.24)]

$$\boldsymbol{\rho} = (\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda} \tag{3.2}$$

With $\mathbf{R}$ available at a central location, the stochastic routing protocols outlined in Section 2.3.2 yield routes maximizing a measure of the arrival rate vector $\boldsymbol{\rho}$. Specifically, letting $f(\boldsymbol{\rho}) : \mathbb{R}^J \to \mathbb{R}$ be a function used to compare arrival rate vectors $\boldsymbol{\rho}$, the optimal routing matrix $\mathbf{T}^*$ is given as the solution of the generic optimization problem:

$$\begin{aligned}(\mathbf{K}^*, \mathbf{T}^*) = \arg\max \;\; & f[(\mathbf{I} - \mathbf{K}_0)\boldsymbol{\lambda}] \\ \text{s.t.} \;\; & K_{ij} = R_{ij}T_{ij} \;\; \text{for } i \neq j, \;\; \mathbf{K}^T\mathbf{1} = \mathbf{1}, \;\; \mathbf{T}^T\mathbf{1} = \mathbf{1}, \;\; \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}.\end{aligned} \tag{3.3}$$

Finding efficient methods to solve (3.3) is challenging for general $f(\boldsymbol{\rho})$. Remarkably though, for any $f(\boldsymbol{\rho})$ that is concave and monotonically non-decreasing in each component Theorem 5 established that the problem in (3.3) can be transformed to an equivalent convex optimization problem for which globally convergent solution methods are available.

In fact, the basic result in Theorem 5 is that for functions that are monotonically non-decreasing per each component there exists an optimal solution of (3.3) with $\boldsymbol{\lambda} = \mathbf{1}$, thus implying that (3.3) can be rewritten as

$$\begin{aligned}(\mathbf{K}^*, \mathbf{T}^*) = \arg\max \;\; & f[(\mathbf{I} - \mathbf{K}_0)\mathbf{1}] \\ \text{s.t.} \;\; & K_{ij} = R_{ij}T_{ij} \;\; \text{for } i \neq j, \;\; \mathbf{K}^T\mathbf{1} = \mathbf{1}, \;\; \mathbf{T}^T\mathbf{1} = \mathbf{1}.\end{aligned} \tag{3.4}$$

The concavity of $f(\boldsymbol{\rho})$ further implies that the argument in (3.4) is concave, which together with the fact that the constraints are linear equalities imply that (3.4) is a convex optimization problem that can be solved in polynomial time using interior point methods [13, Ch. 11].

We emphasize that $\boldsymbol{\lambda} = \mathbf{1}$ is not the unique optimal solution of (3.3) but among the set of optimal pairs $(\mathbf{K}^*, \boldsymbol{\lambda}^*)$ there exists one with $\boldsymbol{\lambda}^* = \mathbf{1}$. In this sense the equivalence of (3.3) and (3.4) does *not* imply that $\boldsymbol{\lambda} = \mathbf{1}$ is necessarily optimal, but that for finding an optimal routing matrix $\mathbf{T}$ it suffices to solve the convex optimization problem in (3.4).

As discussed in Section 2.3.3 requiring $f(\boldsymbol{\rho})$ to be monotonically non-decreasing in each component is a mild condition ensuring that an increase in the rate of one user does not decrease the value of the objective function to be maximized. Many practical rate-maximizing criteria rely on concave functions $f(\boldsymbol{\rho})$ that are monotonically non-decreasing in each component [79]. These include "workhorse criteria" such as optimal $\boldsymbol{\alpha}$-weighted sum-rate with $f(\boldsymbol{\rho}) = \boldsymbol{\alpha}^T \boldsymbol{\rho}$, max-min rate with $f(\boldsymbol{\rho}) = \min_{j\in[1,J]} \rho_j$, and max-product rate criterion with $f(\boldsymbol{\rho}) = \prod_{j\in[1,J]} \rho_j$.

Finding optimal routes as solutions of (3.4) incurs manageable complexity, yet, it requires the reliability matrix $\mathbf{R}$ to be available at the access point (AP - or any designated node for that matter) so that (3.4) can be solved and the optimal routing matrix $\mathbf{T}^*$ can then be distributed through the network. This entails: i) a large communication cost to collect $\mathbf{R}$ and percolate $\mathbf{T}^*$; ii) considerable delay to compute $\mathbf{T}$ in a "batch" mode; and iii) lack of resilience to changes in $\mathbf{R}$, a problem particularly important in highly dynamic (e.g., mobile) scenarios.

Distributed on-line routing algorithms, whereby nodes operate in adaptive mode and iteratively exchange variables only with one-hop neighbors tackle precisely these problems. Indeed, in a distributed iterative algorithm it is assumed that U_j has access only to the link reliabilities for transmission to and from other nodes, i.e., the j-th row and column of $\mathbf{R}$, respectively. Consequently, distributed algorithms neither require $\mathbf{R}$ to be available at a central node, nor percolation of the routing matrix $\mathbf{T}^*$. Thus, they can afford reduced communication cost, and gain robustness to changes in topology due to fading and/or mobility; see e.g., [18, 28, 57, 95, 119].

The main goal of this chapter is to show that the optimization problem in (3.4) can be solved by an iterative distributed algorithm whereby i) node U_j has access only to the j-th row and column of $\mathbf{R}$; ii) U_j interchanges messages with one-hop neighbors, defined

as the set of terminals with positive probability of decoding U_j's packets; and iii) as time progresses U_j computes its optimal routing probabilities, i.e., the j-th column of $\mathbf{T}$.

Notation: For a vector $\mathbf{v} := [v_1, \ldots, v_J]^T$ and a set of indices $c = (i_1, \ldots, i_c)$ with $1 \leq i_1 < \ldots < i_c \leq J$ we define the vector $\mathbf{v}_c := [v_{i_1}, \ldots, v_{i_c}]^T$. Likewise, for the matrix $\mathbf{M} := (M_{ij})$ we define the vectors $\mathbf{M}_{cj} := [M_{i_1 j}, \ldots, M_{i_c j}]^T$ and $\mathbf{M}_{jc} := [M_{j i_1}, \ldots, M_{j i_c}]^T$ containing subsets of the j-th column and row of $\mathbf{M}$ respectively. Note that even if $\mathbf{M}_{jc}$ contains a subset of $\mathbf{M}$'s j-th row it is defined as a column vector. For brevity, the all-zero vector of appropriate dimension is denoted by $\mathbf{0}$, and likewise $\mathbf{1}$ is a vector with all elements equal to one.

3.1 A separable problem

The optimization problem in (3.4) is not in a form that facilitates distributed solution. Towards this end, we first outline in this section equivalent reformulations, whose solution coincides with (3.4) for a given $f(\boldsymbol{\rho})$. The reformulated problems can be separated via a dual decomposition, and lend themselves to distributed solution. For simplicity of exposition let us adopt as optimality criterion the rate of the worst user $f(\boldsymbol{\rho}) = \min_{j \in [1,J]} \rho_j$, leading to the problem

$$(\mathbf{K}^*, \mathbf{T}^*) = \arg\max \min_{j \in [1,J]} [(\mathbf{I} - \mathbf{K}_0)\mathbf{1}]_j$$
$$\text{s.t. } K_{ij} = R_{ij} T_{ij} \text{ for } i \neq j, \quad \mathbf{K}^T \mathbf{1} = \mathbf{1}, \quad \mathbf{T}^T \mathbf{1} = \mathbf{1}. \tag{3.5}$$

In order to reduce the number of variables we will eliminate the equality constraints in (3.5). To this end, define the set $c(j) := \{i : R_{ij} > 0; i \neq j, i \in [1, J+1]\}$ containing the indices of terminals U_i that can decode U_j's transmission with non-zero probability. Likewise, define $r(j) := \{i : R_{ji} > 0; i \neq j, i \in [1, J+1]\}$ as the set of nodes that U_j decodes with non-zero probability. Using these definitions we can write the rate of the j-th user as

$$\rho_j = [(\mathbf{I} - \mathbf{K}_0)\mathbf{1}]_j = 1 - K_{jj} - \sum_{i \in r(j)} K_{ji} = \sum_{i \in c(j)} K_{ij} - \sum_{i \in r(j)} K_{ji} \tag{3.6}$$

where in the second equality we used the constraint $\mathbf{K}^T\mathbf{1} = \mathbf{1}$. Upon substituting $K_{ij} = R_{ij}T_{ij}$, (3.6) becomes

$$\rho_j = \sum_{i \in c(j)} R_{ij}T_{ij} - \sum_{i \in r(j)} R_{ji}T_{ji}. \tag{3.7}$$

For a more compact notation we define the vectors $\bar{\mathbf{s}}_j := \mathbf{T}_{c(j)j}$ and $\bar{\mathbf{s}}'_j = \mathbf{T}_{jc(j)}$ containing the non-zero elements of the j-th column and row of $\mathbf{T}$, respectively. We further define the vectors $\mathbf{r}_j := \mathbf{R}_{c(j)j}$ and $\mathbf{s}_j := \mathbf{R}_{jc(j)}$ so that

$$\rho_j = \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \bar{\mathbf{s}}'_j. \tag{3.8}$$

Vectors $\mathbf{r}_j$, and $\mathbf{s}_j$ are constant and known at node U_j. Indeed, $\mathbf{r}_j := \mathbf{R}_{c(j)j}$ contains the probabilities of other nodes $U_i \neq U_j$ decoding U_j's packets that U_j can easily estimate by counting acknowledgments of packets sent to these terminals. The probabilities of U_j decoding other nodes required to construct $\mathbf{s}_j := \mathbf{R}_{jc(j)}$ can be fed-back from the corresponding (one-hop) neighbors. We assume that estimation of success probabilities and associated feedback among neighboring nodes are perfect.

Using (3.8) and noting that the constraint $\mathbf{T}^T\mathbf{1} = \mathbf{1}$ is equivalent to the set of constraints $\{\bar{\mathbf{s}}_j^T\mathbf{1} = 1\}_{j=1}^J$, we can rewrite the max-min optimal routing problem in (3.5) as

$$\begin{aligned} \mathbf{T}^* = \arg\max \;\; & w \\ \text{s.t.} \;\; & w \preceq \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \bar{\mathbf{s}}'_j = \rho_j, \;\; \bar{\mathbf{s}}_j^T\mathbf{1} = 1, \;\; \mathbf{0} \preceq \bar{\mathbf{s}}_j. \end{aligned} \tag{3.9}$$

Even though (3.9) is written in terms of local variables ($\bar{\mathbf{s}}_j$), local constants ($\mathbf{r}_j$, $\mathbf{s}_j$), and neighboring variables ($\bar{\mathbf{s}}'_j$), it is not yet in a separable form. Indeed, note that i) the variable w is constrained to be smaller than the rates ρ_j of the J terminals and in that sense its optimization requires access to all the variables; and ii) computing ρ_j requires access to the local variables $\bar{\mathbf{s}}_j$ and neighboring variables $\bar{\mathbf{s}}'_j$. While $\bar{\mathbf{s}}_j$ contains U_j's transmission probabilities (the variable that U_j is interested to optimize), $\bar{\mathbf{s}}'_j$ contains the probabilities of other terminals U_i routing their packets through U_j, a variable that U_j's (one-hop) neighbors are interested to optimize.

To overcome these hurdles we introduce local variables w_j and $\mathbf{u}_j$ that we regard as U_j's estimates of (the global variable) w and (the neighboring variable) $\bar{\mathbf{s}}'_j$, and introduce

equality constraints $\mathbf{u}_j = \bar{\mathbf{s}}'_j$ and $w = w_j$, $\forall j \in [1, J]$. Using these (local) variables we can write the constraint in (3.9) as

$$w_j \leq \rho_j = \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \mathbf{u}_j; \quad \mathbf{u}_j = \bar{\mathbf{s}}'_j, \quad w = w_j. \tag{3.10}$$

The last step is to replace w in (3.9) by the weighted sum $w = (\sum_{j=1}^{J} w_j)/J$. Furthermore, if we assume that there is a non-zero probability for a multi-hop route connecting any pair of nodes, the set of constraints $\{w_j = w_i \ \forall i \in c(j)\}_{j=1}^{J}$ is equivalent to requiring $w_i = w_j$, $\forall\, i, j \in [1, J]$. We can now reformulate (3.9) as

$$\begin{aligned} \mathbf{T}^* = \arg\max \ & \frac{1}{J} \sum_{j=1}^{J} w_j \\ \text{s.t.} \ & w_j \preceq \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \mathbf{u}_j, \ \ \bar{\mathbf{s}}_j^T \mathbf{1} = 1, \ \ \mathbf{0} \preceq \bar{\mathbf{s}}_j, \\ & \bar{\mathbf{s}}'_j = \mathbf{u}_j, \ \ w_j = w_i \ \forall i \in c(j) \end{aligned} \tag{3.11}$$

where the maximization is over $\mathbf{T}$, $\{\mathbf{u}_j\}_{j=1}^{J}$, and $\mathbf{w} := [w_1, \ldots, w_J]^T$.

We summarize the equivalence of (3.9) and (3.11) in the following proposition.

Proposition 1 *If there exists a non-zero probability multi-hop route between any pair of nodes the matrix* $\mathbf{T}^*$ *is a solution of (3.9) if and only if it is a solution of (3.11).*

Comparing (3.9) with (3.11) we recognize that the latter does not contain any intrinsically global variable and that the sole coupling between terminals is through the equality constraints $\bar{\mathbf{s}}'_j = \mathbf{u}_j$ and $w_j = w_i$ for all $i \in c(j)$. An important feature of (3.11) is that the constraints on the problem variables can be classified into local constraints involving only variables kept at the j-th terminal and coupling constraints enforcing the equality with neighboring variables of interest. Indeed, note that $w_j \leq \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \mathbf{u}_j$, $\bar{\mathbf{s}}_j^T \mathbf{1} = 1$, and $\mathbf{0} \leq \bar{\mathbf{s}}_j$ involve the variables $\mathbf{x}_j := (w_j, \bar{\mathbf{s}}_j, \mathbf{u}_j)$ only, and can thus be locally enforced, meaning that it is possible for U_j to find values of $\mathbf{x}_j$ satisfying these constraints. The equality constraints cannot be enforced locally but it is important to note that they relate neighboring variables only. Readers familiar with dual decomposition techniques – see, e.g., [9, Sec. 3.4.2] [42, 67] – may notice that the form of (3.11) lends itself to distributed optimization of the type we will elaborate on in Section 3.2.

Remark 5 The terms $\mathbf{r}_j^T \bar{\mathbf{s}}_j$ and $\mathbf{s}_j^T \bar{\mathbf{s}}'_j$ in (3.8) respectively represent the packets successfully transmitted from and to U_j. Their difference is the rate ρ_j available to U_j's own packets. This interpretation of (3.8) is reminiscent of the one encountered in flow control optimization, [9, Sec. 5.1]. Different from flow control, the optimization here over the probabilities T_{ij} has to account for the joint constraint $\bar{\mathbf{s}}_j^T \mathbf{1} = 1$ that outgoing flows from U_j must adhere to. Notwithstanding, flow control is about deterministic splitting of traffic for load balancing purposes, and the optimization of network flows is implemented at the transport layer where optimal routes are assumed available.

3.1.1 Generic problem formulation

The equivalence of (3.5) and (3.11) is not unique to max-min optimal routing since the same steps can be applied to reformulate many optimization problems. To clarify this point consider a given packet success probability matrix $\mathbf{R}$ of which node U_j only knows the non-zero elements of its j-th column and row $\mathbf{r}_j := \mathbf{R}_{c(j)j}$ and $\mathbf{s}_j := \mathbf{R}_{jc(j)}$. Terminal U_j is interested in finding the vector $\bar{\mathbf{s}}_j := \mathbf{T}_{c(j)j}$ that determines its probability of routing packets through neighboring nodes. Introduce the matrix $\mathbf{U}$ with the same sparsity pattern of $\mathbf{T}$ and denote as $\mathbf{u}_j := \mathbf{U}_{jc(j)}$ the non-zero components of the j-th row of $\mathbf{U}$. Node U_j maintains locally the variables w_j, $\bar{\mathbf{s}}_j$, and $\mathbf{u}_j$ that we arrange in the vector $\mathbf{x}_j := [w_j, \bar{\mathbf{s}}_j^T, \mathbf{u}_j^T]^T$. Also, define the vectors $\mathbf{w} := [w_1, \ldots, w_J]^T$ containing the variables w_j of all terminals and $\mathbf{v}_j := \mathbf{w}_{c(j)}$ containing the variables w_j of U_j's neighbors. Finally, abbreviate by $\mathbf{X} := (\mathbf{w}, \mathbf{T}, \mathbf{U})$ the triplet of problem variables.

Our goal in this chapter is to find distributed algorithms converging to the optimal solution of the problem

$$\begin{aligned} \mathbf{X}^* := (\mathbf{T}^*, \mathbf{U}^*, \mathbf{w}^*) = \arg\max_{\mathbf{X}} \quad & \mathbf{w}^T \mathbf{1} \\ \text{s.t.} \quad & \mathbf{x}_j := (w_j, \bar{\mathbf{s}}_j, \mathbf{u}_j) \in \mathcal{X}_j, \quad \bar{\mathbf{s}}'_j = \mathbf{u}_j, \quad \mathbf{v}_j = w_j \mathbf{1} \end{aligned} \tag{3.12}$$

where $\mathcal{X}_j$ is a set that defines the specific routing optimality criteria. Note that the constraint $\bar{\mathbf{s}}'_j = \mathbf{u}_j$ implies that $\mathbf{T}^* = \mathbf{U}^*$ so that after obtaining the optimal solution U_j knows the (optimal) probabilities $\bar{\mathbf{s}}_j$ with which to route its packets through its neighbors and the

probabilities $\bar{\mathbf{s}}'_j = \mathbf{u}_j$ with which its neighbors route packets through himself.

To find the optimal solution to (3.12) we require the following assumptions to hold true:

(a1) The set $\mathcal{X}_j$ is convex.

(a2) There is a non-zero probability multi-hop route connecting any pair of nodes.

(a3) Node U_j can communicate with its one hop neighbors $\{U_i : i \in c(j)\}$ (it does not have access to variables of other nodes).

(a4) The probability that U_j decodes U_i is non-zero if and only if the probability that U_i decodes U_j is non-zero. This implies $c(j) = r(j)$ for all $j \in [1, J]$.

Assumption (a1) ensures that the problem in (3.12) is convex; (a2) is required so that the constraints $\{w_j = w_i \ \forall i \in c(j)\}_{j=1}^{J}$ imply $w_i = w_j$, $\forall i, j \in [1, J]$; (a3) is in line with the distributed setup; and (a4) guarantees that if U_j has access to U_i's variables then U_i has access to U_j's variables, which is natural in a peer-to-peer setting, and will be exploited later on.

The formulation in (3.12) encompasses all the routing problems defined in [79], with the set $\mathcal{X}_j$ specifying the corresponding optimality criterion. In particular we have:

Max-min optimal rate. This is the problem considered in detail in Section 3.1 and can be obtained from (3.12) by defining the set

$$\mathcal{X}_j^1 = \left\{\mathbf{x}_j : w_j \leq \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \mathbf{u}_j, \quad \mathbf{0} \preceq \bar{\mathbf{s}}_j, \quad \bar{\mathbf{s}}_j^T \mathbf{1} = 1\right\}. \tag{3.13}$$

Additional convex constraints can be added to the definition of $\mathcal{X}_j$. Since we know that $\mathbf{u}_j$ is a vector of probabilities using the set $\mathcal{X}_j = \mathcal{X}_j^1 \cap \{\mathbf{x}_j : \mathbf{0} \preceq \mathbf{u}_j \preceq \mathbf{1}\}$ is equivalent to using $\mathcal{X}_j^1$ because the constraint $\mathbf{0} \preceq \mathbf{u}_j \preceq \mathbf{1}$ is implicit in $\mathbf{u}_j = \bar{\mathbf{s}}'_j$. Preventing the components of $\mathbf{u}_j$ to become too large contributes to the numerical stability of the problem.

Optimal weighted sum-rate. We want to maximize a weighted sum of the rates, i.e., $f(\boldsymbol{\rho}) = \boldsymbol{\alpha}^T \boldsymbol{\rho}$ with $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_J]^T \succeq \mathbf{0}$. In this case we define the set

$$\mathcal{X}_j^2 := \left\{\mathbf{x}_j : w_j = \alpha_j(\mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \mathbf{u}_j), \quad \mathbf{0} \preceq \bar{\mathbf{s}}_j, \quad \bar{\mathbf{s}}_j^T \mathbf{1} = 1\right\} \tag{3.14}$$

and consider the optimization problem

$$\begin{aligned} \mathbf{T}^* = \arg\max \ & \mathbf{w}^T\mathbf{1} \\ \text{s.t.} \ & \mathbf{x}_j := (w_j, \bar{\mathbf{s}}_j, \mathbf{u}_j) \in \mathcal{X}_j^2, \ \ \bar{\mathbf{s}}_j' = \mathbf{u}_j \end{aligned} \tag{3.15}$$

which amounts to dropping the constraint $\mathbf{v}_j = w_j\mathbf{1}$ in (3.12). Note that for this criterion $w_j = \alpha_j\rho_j$.

Extra convex constraints can be dealt with by modifying the set $\mathcal{X}_j^2$ as in the previous example. A case of interest is to consider a minimum acceptable rate $\rho_j^{\min}$ for terminal U_j that can be managed by considering the set $\mathcal{X}_j := \mathcal{X}_j^2 \cap \{w_j/\alpha_j \geq \rho_j^{\min}\}$. A solution $\mathbf{T}^*$ to (3.15) with a minimum acceptable rate constraint may not exist for some values of $\boldsymbol{\rho}^{\min}$ – in such cases interior point methods return an infeasibility certificate. When it exists, $\mathbf{T}^*$ ensures the minimum acceptable rate $\rho_j^{\min}$ to every user with the excess traffic distributed to the most favored users with large values of α_j and/or reliable connections to one of the APs.

Optimal product of rates. Maximizing the product of rates constitutes a more fair alternative to the maximum sum-rate criterion in (3.15) since it prevents solutions in which some users receive a very small packet delivery rate. The function to be maximized in this case is $f(\boldsymbol{\rho}) = \prod_{j=1}^J \rho_j$. Equivalently, since the logarithm is monotonically increasing the concave function $f(\boldsymbol{\rho}) = \sum_{j=1}^J \log(\rho_j)$ can be used instead. To cast this problem under the distributable formulation in (3.12) it suffices to define the set

$$\mathcal{X}_j^3 := \left\{\mathbf{x}_j : w_j \leq \log[\mathbf{r}_j^T\bar{\mathbf{s}}_j - \mathbf{s}_j^T\mathbf{u}_j], \ \ \mathbf{0} \preceq \bar{\mathbf{s}}_j, \ \ \bar{\mathbf{s}}_j^T\mathbf{1} = 1\right\} \tag{3.16}$$

and replace $\mathcal{X}_j^2$ by $\mathcal{X}_j^3$ in (3.15). The local components of the argument w_j denote the logarithm of the local rate.

Another example of a convex constraint is a cooperation limit whereby terminals require their own rate to be at least a certain percentage $\beta_j \in [0, 1]$ of their total outgoing rate $\mathbf{r}_j^T\bar{\mathbf{s}}_j$. To add this constraint define the set $\mathcal{X}_j := \mathcal{X}_j^3 \cap \{\mathbf{x}_j : \mathbf{r}_j^T\bar{\mathbf{s}}_j - \mathbf{s}_j^T\mathbf{u}_j \geq \beta_j(\mathbf{r}_j^T\bar{\mathbf{s}}_j)\}$. This constraint guarantees that at least β_j of the packets that U_j transmits were generated at U_j.

Optimal rate with relays. In a relay network a group of terminals collaborate in relaying traffic on behalf of a designated active user. Let U_{j_0} denote this active user and terminals $\{U_j\}_{j=1,j\neq j_0}^{J}$ be the relays. The optimal relay network maximizing the rate ρ_{j_0} can be found by solving (3.12) with

$$\begin{aligned}\mathcal{X}_j^3 &= \left\{\mathbf{x}_j : 0 = \mathbf{r}_j^T\bar{\mathbf{s}}_j - \mathbf{s}_j^T\mathbf{u}_j,\ \ \mathbf{0} \preceq \bar{\mathbf{s}}_j,\ \ \bar{\mathbf{s}}_j^T\mathbf{1} = 1\right\},\ j \neq j_0\\ \mathcal{X}_{j_0}^3 &= \left\{\mathbf{x}_{j_0} : w_{j_0} = \mathbf{r}_{j_0}^T\bar{\mathbf{s}}_{j_0} - \mathbf{s}_{j_0}^T\mathbf{u}_{j_0},\ \ \mathbf{0} \preceq \bar{\mathbf{s}}_{j_0},\ \ \bar{\mathbf{s}}_{j_0}^T\mathbf{1} = 1\right\}.\end{aligned} \tag{3.17}$$

In this example, w_j is the local estimate of the source's rate ρ_{j_0} at terminal U_j.

3.2 Distributed implementation via dual decomposition

Problems of the form (3.12) or (3.15) can be solved using the so called dual decomposition methods [9, Sec. 3.4.2], [67]. Since (a1) guarantees convexity of the problem the basic idea is to optimize the dual function that, as we will show in this section, exhibits a separable structure. Associate, thus, Lagrange multipliers $\boldsymbol{\lambda}_j$ with the constraints $\bar{\mathbf{s}}_j' - \mathbf{u}_j = \mathbf{0}$ and $\boldsymbol{\mu}_j$ with the constraints $\mathbf{v}_j - w_j\mathbf{1} = \mathbf{0}$ to form the Lagrangian

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{M}) = -\mathbf{w}^T\mathbf{1} + \sum_{j=1}^{J}\left[(\bar{\mathbf{s}}_j' - \mathbf{u}_j)^T\boldsymbol{\lambda}_j + (\mathbf{v}_j - w_j\mathbf{1})^T\boldsymbol{\mu}_j\right] \tag{3.18}$$

which is defined over the feasible region of the primal variables $\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^{J}$. Matrices $\boldsymbol{\Lambda}$ and $\mathbf{M}$ are defined to have the same sparsity pattern of $\mathbf{T}$ (and thus $\mathbf{U}$); the dual variables (multipliers) in (3.18) are respectively given by the non-zero elements of the j-th row of $\boldsymbol{\Lambda}$, and the j-th column of $\mathbf{M}$; i.e., $\boldsymbol{\lambda}_j = \boldsymbol{\Lambda}_{jc(j)}$ and $\boldsymbol{\mu}_j = \mathbf{M}_{c(j)j}$. We assume that $\boldsymbol{\lambda}_j$ and $\boldsymbol{\mu}_j$ are kept by terminal U_j.

The Lagrangian in (3.18) is used to obtain the dual function

$$g(\boldsymbol{\Lambda}, \mathbf{M}) = \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^{J}} \mathcal{L}(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{M}) \tag{3.19}$$

which in turn leads to the dual problem defined as the unconstrained maximization of $g(\boldsymbol{\Lambda}, \mathbf{M})$ – note that we do not impose non-negativity constraints on the multipliers because this is an equality-constrained problem. For convex optimization problems strong duality

holds, implying that the maximum in (3.12) coincides with the negative of the maximum of $g(\mathbf{\Lambda}, \mathbf{M})$; i.e.,

$$\mathbf{1}^T \mathbf{w}^* = -\max_{\mathbf{\Lambda}, \mathbf{M}} g(\mathbf{\Lambda}, \mathbf{M}). \tag{3.20}$$

The problem in (3.20) is an unconstrained optimization problem that can be solved with a gradient ascent algorithm. However, since the dual function $g(\mathbf{\Lambda}, \mathbf{M})$ is not always differentiable a generalization of the gradient, the so called *subgradient* is used instead.

Definition 2 *Consider a concave function $f(\mathbf{\Lambda}) : \mathbb{R}^M \to \mathbb{R}$. If $\nabla_{\mathbf{\Lambda}}(\mathbf{\Lambda})$ satisfies*

$$f(\tilde{\mathbf{\Lambda}}) \leq f(\mathbf{\Lambda}) + \nabla_{\mathbf{\Lambda}}(\mathbf{\Lambda})(\tilde{\mathbf{\Lambda}} - \mathbf{\Lambda}) \tag{3.21}$$

for all $\tilde{\mathbf{\Lambda}} \in \mathbb{R}^M$ we say that $\nabla_{\mathbf{\Lambda}}(\mathbf{\Lambda})$ is a subgradient of $f(\mathbf{\Lambda})$ at $\mathbf{\Lambda}$. Given a subset of components $\boldsymbol{\lambda}$ of $\mathbf{\Lambda}$ we denote as $\nabla_{\boldsymbol{\lambda}}(\mathbf{\Lambda})$ the corresponding components of $\nabla_{\mathbf{\Lambda}}(\mathbf{\Lambda})$.

The subgradient is any vector $\nabla_{\mathbf{\Lambda}}$ defining a supporting hyperplane of the concave function $f(\mathbf{\Lambda})$. When a gradient exists, i.e., when $f(\mathbf{\Lambda})$ is differentiable, it is the unique subgradient of $f(\mathbf{\Lambda})$. A subgradient of $g(\mathbf{\Lambda}, \mathbf{M})$ is presented in the next proposition [9, Sec. 3.4.2].

Proposition 2 *For given multipliers $\mathbf{\Lambda}$ and $\mathbf{M}$, let $\mathbf{X}^\dagger(\mathbf{\Lambda}, \mathbf{M})$ denote the optimal argument of the Lagrangian, i.e.,*

$$\mathbf{X}^\dagger(\mathbf{\Lambda}, \mathbf{M}) := \arg \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) \tag{3.22}$$

with $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M})$ given by (3.18). Then, a subgradient $\nabla_{\mathbf{\Lambda},\mathbf{M}}$ of $g(\mathbf{\Lambda}, \mathbf{M})$ has components,

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}_j}(\mathbf{\Lambda}, \mathbf{M}) &= \bar{\mathbf{s}}_j'^\dagger(\mathbf{\Lambda}, \mathbf{M}) - \mathbf{u}_j^\dagger(\mathbf{\Lambda}, \mathbf{M}) \\ \nabla_{\boldsymbol{\mu}_j}(\mathbf{\Lambda}, \mathbf{M}) &= \mathbf{v}_j^\dagger(\mathbf{\Lambda}, \mathbf{M}) - w_j^\dagger(\mathbf{\Lambda}, \mathbf{M})\mathbf{1}. \end{aligned} \tag{3.23}$$

Proof: Consider the value of the dual function at an arbitrary point $(\tilde{\mathbf{\Lambda}}, \tilde{\mathbf{M}})$

$$g(\tilde{\mathbf{\Lambda}}, \tilde{\mathbf{M}}) = \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} -\mathbf{w}^T\mathbf{1} + \sum_{j=1}^J \left[(\bar{\mathbf{s}}_j' - \mathbf{u}_j)^T \tilde{\boldsymbol{\lambda}}_j + (\mathbf{v}_j - w_j\mathbf{1})^T \tilde{\boldsymbol{\mu}}_j \right] \tag{3.24}$$

$$\leq -\mathbf{w}^{\dagger T}\mathbf{1} + \sum_{j=1}^J \left[(\bar{\mathbf{s}}_j'^\dagger - \mathbf{u}_j^\dagger)^T \tilde{\boldsymbol{\lambda}}_j + (\mathbf{v}_j^\dagger - w_j^\dagger\mathbf{1})^T \tilde{\boldsymbol{\mu}}_j \right] \tag{3.25}$$

where for notational simplicity we omit the arguments of $\bar{\mathbf{s}}_j'^{\dagger}$, $\mathbf{u}^{\dagger}$, $\mathbf{v}_j^{\dagger}$, and $w_j^{\dagger}$. The equality in (3.24) follows from the definitions of the dual function in (3.19) and the Lagrangian in (3.18); and the inequality in (3.25) is true since $\{\mathbf{x}^{\dagger}\}_{j=1}^{J} = \mathbf{X}^{\dagger}$ cannot yield a value smaller than the optimal argument of (3.24).

Subtracting $g(\mathbf{\Lambda}, \mathbf{M}) = -\mathbf{w}^{\dagger T}\mathbf{1} + \sum_{j=1}^{J}\Big[(\bar{\mathbf{s}}_j'^{\dagger} - \mathbf{u}_j^{\dagger})^T\boldsymbol{\lambda}_j + (\mathbf{v}_j^{\dagger} - w_j^{\dagger}\mathbf{1})^T\boldsymbol{\mu}_j\Big]$ from both sides of the inequality in (3.25) yields

$$g(\tilde{\mathbf{\Lambda}}, \tilde{\mathbf{M}}) - g(\mathbf{\Lambda}, \mathbf{M}) \leq \sum_{j=1}^{J}\Big[(\bar{\mathbf{s}}_j'^{\dagger} - \mathbf{u}_j^{\dagger})^T(\tilde{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j) + (\mathbf{v}_j^{\dagger} - w_j^{\dagger}\mathbf{1})^T(\tilde{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j)\Big] \tag{3.26}$$

Comparing (3.26) with (3.21) we see that the constraint violations in (3.23) satisfy the definition of a subgradient of $g(\mathbf{\Lambda}, \mathbf{M})$ [cf. (3.21)]. □

Proposition 2 tell us that for general multipliers $(\mathbf{\Lambda}, \mathbf{M})$ the Lagrangian is optimized by variables $\mathbf{X}^{\dagger}(\mathbf{\Lambda}, \mathbf{M})$ that violate the equality constraints in (3.12). Interestingly, the amount by which the equality constraints are violated is a subgradient of the dual function. Indeed, the multiplier $\boldsymbol{\lambda}_j$ ($\boldsymbol{\mu}_j$) is associated with the constraint $\bar{\mathbf{s}}_j' - \mathbf{u}_j = 0$ ($\mathbf{v}_j - w_j\mathbf{1} = 0$); the optimal arguments of the Lagrangian violate this constraint by an amount $\nabla_{\boldsymbol{\lambda}_j}(\mathbf{\Lambda}, \mathbf{M}) = \bar{\mathbf{s}}_j'^{\dagger}(\mathbf{\Lambda}, \mathbf{M}) - \mathbf{u}_j^{\dagger}(\mathbf{\Lambda}, \mathbf{M})$ ($\nabla_{\boldsymbol{\mu}_j}(\mathbf{\Lambda}, \mathbf{M}) = \mathbf{v}_j^{\dagger}(\mathbf{\Lambda}, \mathbf{M}) - w_j^{\dagger}(\mathbf{\Lambda}, \mathbf{M})\mathbf{1}$).

A important property of the optimal arguments of the Lagrangian is that they can be computed locally at each node. To be precise define the vectors $\boldsymbol{\lambda}_j' = \mathbf{\Lambda}_{c(j)j}$ and $\boldsymbol{\mu}_j' = \mathbf{M}_{jc(j)}$ containing the dual variables of the one hop neighbors $\{U_i : i \in c(j)\}$, and construct the local Lagrangian $\mathcal{L}_j(\mathbf{x}_j; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j', \boldsymbol{\mu}_j')$ by grouping the terms that depend only on the local variable $\mathbf{x}_j$ [cf. (3.18)]

$$\mathcal{L}_j(\mathbf{x}_j; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j', \boldsymbol{\mu}_j') = -w_j + \bar{\mathbf{s}}_j^T\boldsymbol{\lambda}_j' - \mathbf{u}_j^T\boldsymbol{\lambda}_j + w_j\mathbf{1}^T(\boldsymbol{\mu}_j' - \boldsymbol{\mu}_j). \tag{3.27}$$

By construction $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) = \sum_{j=1}^{J}\mathcal{L}_j(\mathbf{x}_j; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j', \boldsymbol{\mu}_j')$ [cf. (3.18) and (3.27)]. If we further note that the primal variables $\mathbf{x}_j$ appear only in $\mathcal{L}_j(\mathbf{x}_j; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j', \boldsymbol{\mu}_j')$ we conclude that the optimal arguments in (3.22) can be found as

$$\mathbf{x}_j^{\dagger} := \arg\min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{L}_j(\mathbf{x}_j, \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j', \boldsymbol{\mu}_j'). \tag{3.28}$$

The ultimate reasons enabling a distributed implementation of a subgradient ascent algorithm can be read out from Proposition 2 and (3.28): i) a subgradient of the dual function

is obtained from the arguments optimizing the Lagrangian $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M})$ [cf. (3.23)]; ii) the subgradients $\nabla_{\boldsymbol{\lambda}_j}(\mathbf{\Lambda}, \mathbf{M})$ and $\nabla_{\boldsymbol{\mu}_j}(\mathbf{\Lambda}, \mathbf{M})$ depend only on local and neighboring variables [cf. (3.23)]; and iii) the optimization of the Lagrangian $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M})$ separates into the optimization of J local Lagrangians $\mathcal{L}_j(\mathbf{x}_j, \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\lambda}'_j, \boldsymbol{\mu}'_j)$, furthermore, these local Lagrangians depend only on local and neighboring variables [cf. (3.27) and (3.28)].

Consequently, subgradient ascent for $g(\mathbf{\Lambda}, \mathbf{M})$ can be implemented by the following distributable iteration:

[I1] *Compute subgradient.* Given local multipliers $\boldsymbol{\lambda}_j(n)$ and $\boldsymbol{\mu}_j(n)$, and neighboring multipliers $\boldsymbol{\lambda}'_j(n)$ and $\boldsymbol{\mu}'_j(n)$, minimize the local Lagrangian with respect to the local primal variables,

$$\mathbf{x}_j(n) = \arg \min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{L}_j[\mathbf{x}_j, \boldsymbol{\lambda}_j(n), \boldsymbol{\mu}_j(n), \boldsymbol{\lambda}'_j(n), \boldsymbol{\mu}'_j(n)] := \arg \min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{L}_j(\mathbf{x}_j, n) \tag{3.29}$$

where we defined $\mathcal{L}_j(\mathbf{x}_j, n) := \mathcal{L}_j[\mathbf{x}_j, \boldsymbol{\lambda}_j(n), \boldsymbol{\mu}_j(n), \boldsymbol{\lambda}'_j(n), \boldsymbol{\mu}'_j(n)]$ and the primal iterates are $\mathbf{x}_j(n) := [w_j(n), \bar{\mathbf{s}}_j(n), \mathbf{u}_j(n)]$.

[I2] *Subgradient ascent step.* Using local primal variables $[w_j(n), \bar{\mathbf{s}}_j(n), \mathbf{u}_j(n)]$ and neighboring primal variables $[\mathbf{v}_j(n), \bar{\mathbf{s}}'_j(n), \mathbf{u}'_j(n)]$ update local multipliers

$$\begin{aligned} \boldsymbol{\lambda}_j(n+1) &= \boldsymbol{\lambda}_j(n) + c_n[\bar{\mathbf{s}}'_j(n) - \mathbf{u}_j(n)] \\ \boldsymbol{\mu}_j(n+1) &= \boldsymbol{\mu}_j(n) + c_n[\mathbf{v}_j(n) - w_j(n)\mathbf{1}] \end{aligned} \tag{3.30}$$

where c_n is a properly selected step size.

Algorithm 2 details the distributed implementation of [I1]-[I2]. Given the local multipliers $\boldsymbol{\lambda}_j(n)$ and $\boldsymbol{\mu}_j(n)$, and the one-hop-neighbors' multipliers $\boldsymbol{\lambda}'_j(n)$ and $\boldsymbol{\mu}'_j(n)$, user terminal U_j solves a (local) convex optimization problem to find the primal variables $\mathbf{x}_j(n)$ that optimize the (local and global) Lagrangian; step 3. In turn, these primal variables are used in the gradient ascent steps 6 and 7 to obtain the updated multipliers $\boldsymbol{\lambda}_j(n+1)$ and $\boldsymbol{\mu}_j(n+1)$. Steps 6 and 7 represent the subgradient ascent step for the dual function $g(\mathbf{\Lambda}, \mathbf{M})$ and as such are the steps guaranteeing convergence of the iterates $\{\boldsymbol{\lambda}_j(n)\}_{j=1}^J$ and $\{\boldsymbol{\mu}_j(n)\}_{j=1}^J$ obtained from (3.29)-(3.30) to $\{\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*\}_{j=1}^J := \arg\max g(\mathbf{\Lambda}, \mathbf{M})$ as $n \to \infty$ (convergence

Algorithm 2 Dual decomposition solver

Require: Packet success probabilities to and from neighbors $\mathbf{R}_{c(j)j}$ and $\mathbf{R}_{jc(j)}$

Ensure: Optimal multipliers $\boldsymbol{\lambda}_j^*$ and $\boldsymbol{\mu}_j^*$

1: **for** $n = 1$ to ∞ **do** {repeat for the life of the network}

2: Receive multipliers $\lambda_{ij}(n)$ and $\mu_{ji}(n)$ from one hop neighbors $\{U_i : i \in c(j)\}$

3: Minimize Lagrangian [cf. (3.29)]: $\mathbf{x}_j(n) = \arg\min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{L}_j[\mathbf{x}_j, \boldsymbol{\lambda}_j(n), \boldsymbol{\mu}_j(n), \boldsymbol{\lambda}_j'(n), \boldsymbol{\mu}_j'(n)]$

4: Transmit $w_j(n)$, $t_{ij}(n)$, and $u_{ji}(n)$ to neighbor U_i; repeat for all $\{U_i : i \in c(j)\}$.

5: Receive $w_i(n)$, $t_{ji}(n)$, and $u_{ij}(n)$ from one hop neighbors $\{U_i : i \in c(j)\}$

6: Subgradient ascent iteration for $\boldsymbol{\lambda}_j$ [cf. (3.30)]: $\boldsymbol{\lambda}_j(n+1) = \boldsymbol{\lambda}_j(n) + c_n[\bar{\mathbf{s}}_j'(n) - \mathbf{u}_j(n)]$

7: Subgradient ascent iteration for $\boldsymbol{\mu}_j$ [cf. (3.30)]: $\boldsymbol{\mu}_j(n+1) = \boldsymbol{\mu}_j(n) + c_n[\mathbf{v}_j(n) - w_j(n)\mathbf{1}]$

8: Transmit multipliers $\lambda_{ji}(n+1)$ and $\mu_{ij}(n+1)$ to neighbor U_i; repeat for all $\{U_i : i \in c(j)\}$

9: **end for**

of (3.29)-(3.30) requires some qualifications that we discuss in the next subsection). The remaining steps ensure that the variables are properly communicated. Steps 8 and 2 ensure that the updated multipliers are sent to and received by the corresponding neighboring node, while steps 4 and 5 guarantee the same for the primal variables.

3.2.1 Discussion of convergence properties

The goal of Algorithm 2 is for U_j to obtain the optimal routing probabilities $\bar{\mathbf{s}}_j$. We are thus interested in having $\lim_{n\to\infty} \bar{\mathbf{s}}_j(n) = \bar{\mathbf{s}}_j^*$, with $\bar{\mathbf{s}}_j(n)$ obtained from the iteration (3.29)-(3.30) and $\bar{\mathbf{s}}_j^*$ the solution of (3.12). Since the iteration (3.29)-(3.30) implements subgradient ascent for the dual function, convergence of the primal variables cannot be always guaranteed. Relevant convergence properties of subgradient descent are summarized next (see e.g., [9, Sec. 3.4.3] and [18]).

Property 1 *Consider the iteration (3.29)-(3.30) and let* $\{\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*\}_{j=1}^J := \arg\max g(\boldsymbol{\Lambda}, \mathbf{M})$ *denote the optimal solution of the dual problem in (3.20). We then have that*

(a) if the step size is constant, i.e., $c_n = c\ \forall n$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{n=0}^{\infty} \boldsymbol{\lambda}_j(n) = \boldsymbol{\lambda}_j^* \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{n=0}^{\infty} \boldsymbol{\mu}_j(n) = \boldsymbol{\mu}_j^* \tag{3.31}$$

implying that the average value of the dual iterates converges to the optimal dual variables; and

(b) if the step size sequence is non-summable, $\sum_{n=0}^{\infty} c_n = \infty$ but square summable, $\sum_{n=0}^{\infty} c_n^2 < \infty$

$$\lim_{n\to\infty} \boldsymbol{\lambda}_j(n) = \boldsymbol{\lambda}_j^* \qquad \lim_{n\to\infty} \boldsymbol{\mu}_j(n) = \boldsymbol{\mu}_j^* \tag{3.32}$$

implying that the sequence of dual iterates converges to the optimal dual variables.

Convergence of the dual iterates in the sense described in Property 1, does not imply that the same holds true for $\mathbf{x}_j$, and in practice $\lim_{n\to\infty} \bar{\mathbf{s}}_j(n) \neq \bar{\mathbf{s}}_j^*$ for many practical optimality criteria. This is particularly true when (3.12) amounts to a linear program, a class that includes max-min optimal rate, optimal weighted sum-rate, and optimal rate with relays as defined in Section 3.1.1 – for these problems the sets $\mathcal{X}_j$ are convex polygons. Many regularization approaches are known to guarantee convergence of the primal iterates $\mathbf{x}_j(n)$ to the primal optima $\mathbf{x}_j^*$. One of them, the method of multipliers, is discussed in the next section.

Remark 6 Avoiding a large variance of the iterates (i.e., large fluctuations around the mean) when $c_n = c\ \forall n$ as in Property 1 - (a) requires a small value of c. However, this entails a slow convergence rate. This can be alleviated by adjusting c_n as per Property 1 - (b), but this is difficult to implement in a distributed setting. These complementary drawbacks provide another motivation for the approach in Section 3.3.

3.3 The method of multipliers

While useful as a first approach, the dual decomposition method summarized in Algorithm 2 does not always lead to a satisfactory solution of (3.12). As discussed previously, when the

dual function is non-differentiable and the step size c_n is fixed (3.29) - (3.30) converges only on an average sense. Perhaps more important, recovering the primal variables optimizing (3.12) from the dual variables optimizing (3.20) cannot always be guaranteed.

A common regularization approach is the so called method of multipliers (MOM). The MOM is based on modifying the optimization argument in (3.12) by adding (hence the term *regularization*) a quadratic term corresponding to the squared norm of the equality constraints,

$$\begin{aligned}\mathbf{T}^* = \arg\min_{\mathbf{X}} \quad & -\mathbf{w}^T\mathbf{1} + \frac{c}{2}\sum_{j=1}^{J}\left[\|\bar{\mathbf{s}}'_j - \mathbf{u}_j\|^2 + \|\mathbf{v}_j - w_j\mathbf{1})\|^2\right] \\ \text{s.t.} \quad & \mathbf{x}_j := (w_j, \bar{\mathbf{s}}_j, \mathbf{u}_j) \in \mathcal{X}_j; \;\; \bar{\mathbf{s}}'_j = \mathbf{u}_j; \;\; \mathbf{v}_j = w_j\mathbf{1}. \end{aligned} \tag{3.33}$$

Due to the triangle inequality, norms are convex functions of their arguments, and consequently the problem in (3.33) is convex. Furthermore, the solutions of (3.12) and (3.33) coincide since the terms $\|\bar{\mathbf{s}}'_j - \mathbf{u}_j\|^2$ and $\|\mathbf{v}_j - w_j\mathbf{1}\|^2$ are null at any feasible point. The Lagrangian associated with (3.33) is known as the *augmented* Lagrangian of (3.12) and is given by

$$\mathcal{A}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) = \mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) + \frac{c}{2}\sum_{j=1}^{J}\left[\|\bar{\mathbf{s}}'_j - \mathbf{u}_j\|^2 + \|\mathbf{v}_j - w_j\mathbf{1})\|^2\right] \tag{3.34}$$

with $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M})$ as in (3.18).

Mimicking steps (3.19) and (3.20) we can define the dual function $h(\mathbf{\Lambda}, \mathbf{M}) := \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M})$ and conclude that finding the optimal value of (3.33) – which coincides with the optimal value of (3.12) – is equivalent to solving the corresponding dual problem

$$-\mathbf{1}^T\mathbf{w}^* \;=\; \max_{\mathbf{\Lambda},\mathbf{M}} h(\mathbf{\Lambda}, \mathbf{M}) \;:=\; \max_{\mathbf{\Lambda},\mathbf{M}} \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) \tag{3.35}$$

Recalling Proposition 2 we can obtain a subgradient of $h(\mathbf{\Lambda}, \mathbf{M})$ from the arguments minimizing the augmented Lagrangian; (re-) defining

$$\begin{aligned}\{\mathbf{x}_j^\dagger(\mathbf{\Lambda}, \mathbf{M})\}_{j=1}^J &= \arg\min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) \\ &= \arg\min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{L}(\mathbf{X}, \mathbf{\Lambda}, \mathbf{M}) + \frac{c}{2}\sum_{j=1}^{J}\left[\|\bar{\mathbf{s}}'_j - \mathbf{u}_j\|^2 + \|\mathbf{v}_j - w_j\mathbf{1})\|^2\right] \end{aligned} \tag{3.36}$$

we have that the subgradient components $\nabla_{\boldsymbol{\lambda}_j}(\boldsymbol{\Lambda}, \mathbf{M})$ and $\nabla_{\boldsymbol{\mu}_j}(\boldsymbol{\Lambda}, \mathbf{M})$ of $h(\boldsymbol{\Lambda}, \mathbf{M})$ are given as in (3.23).

Recapitulating the fundamental motives leading to a distributed implementation of subgradient ascent for $g(\boldsymbol{\Lambda}, \mathbf{M})$ we see that: i) the arguments minimizing the augmented Lagrangian $\mathcal{A}(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{M})$ lead to a subgradient of $h(\boldsymbol{\Lambda}, \mathbf{M})$ [cf. (3.23) and (3.36)]; and ii) the subgradients $\nabla_{\boldsymbol{\lambda}_j}(\boldsymbol{\Lambda}, \mathbf{M})$ and $\nabla_{\boldsymbol{\mu}_j}(\boldsymbol{\Lambda}, \mathbf{M})$ depend only on local and neighboring variables [cf. (3.23)].

Unlike the case of $\mathcal{L}(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{M})$, the minimization of $\mathcal{A}(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{M})$ cannot be separated into local independent optimizations due to coupling between $\bar{\mathbf{s}}'_j$ and $\mathbf{u}_j$ and $\mathbf{v}_j$ and w_j introduced by the quadratic terms [cf. (3.36)]. Note, however, that the coupling is between neighboring variables only, and consequently we can - again - devise a distributed algorithm to solve the minimization in (3.36). To be precise, our goal is a distributed algorithm that for given multipliers $\boldsymbol{\Lambda}(n)$ and $\mathbf{M}(n)$ at the n-th iteration converges to the optimal value of the augmented Lagrangian

$$\{\mathbf{x}_j(n)\}_{j=1}^J := \arg \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, \boldsymbol{\Lambda}(n), \mathbf{M}(n)) := \arg \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, n) \tag{3.37}$$

A separable iteration converging to $\{\mathbf{x}_j(n)\}_{j=1}^J$ can be obtained using coordinate descent as described in the following proposition.

Proposition 3 *For fixed n consider iterations in a second index m. With $\mathcal{L}_j(\mathbf{x}_j, n)$ as in (3.29) define the local augmented Lagrangian at the (n, m)-th iteration as*

$$\mathcal{A}_j(\mathbf{x}_j, n, m) = \mathcal{L}_j(\mathbf{x}_j, n) + \frac{c}{2}\left[\|\bar{\mathbf{s}}'_j(n, m) - \mathbf{u}_j\|^2 + 2\|\mathbf{v}_j(n, m) - w_j\mathbf{1})\|\|^2 + \|\bar{\mathbf{s}}_j - \mathbf{u}'_j(n, m)\|^2\right] \tag{3.38}$$

and consider iterates $\{\mathbf{x}_j(n, m+1)\}_{j=1}^J$ satisfying

$$\mathbf{x}_j(n, m+1) = \arg \min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{A}_j(n, m). \tag{3.39}$$

Then, $\mathbf{x}_j(n, m)$ converges to the optimal value of (3.37), i.e., $\lim_{m\to\infty} \mathbf{x}_j(n, m) = \mathbf{x}_j(n)$, $\forall$ $j \in [1, J]$.

Proof: The only terms of the augmented Lagrangian $\mathcal{A}(\mathbf{X}, n)$ in (3.37) that depend on local primal variables $\mathbf{x}_j$ are those contained in the local augmented lagrangian $\mathcal{A}_j(\mathbf{x}_j, n, m)$ in (3.38). Thus, the optimal arguments in (3.39) are such that

$$\mathbf{x}_j(n, m+1) = \arg \min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{A}[\mathbf{x}_1(n,m), \ldots, \mathbf{x}_{j-1}(n,m), \mathbf{x}_j, \mathbf{x}_{j+1}(n,m), \ldots, \mathbf{x}_J(n,m), n]. \tag{3.40}$$

That is, (3.40) minimizes $\mathcal{A}(\mathbf{X}, n)$ along the coordinates corresponding to $\mathbf{x}_j$. By definition this is a coordinate descent algorithm for minimizing $\mathcal{A}(\mathbf{X}, n)$ and we thus have

$$\lim_{m \to \infty} \mathbf{x}_j(n, m) = \arg \min_{\{\mathbf{x}_j \in \mathcal{X}_j\}_{j=1}^J} \mathcal{A}(\mathbf{X}, n) =: \mathbf{x}_j(n). \tag{3.41}$$

where the first equality follows from convergence results for coordinate descent, see e.g., [9, Sec. 3.2.1]; and the second one from the definition in (3.37). □

The coordinate descent iteration (3.38)-(3.39) depends only on local and neighboring variables and it can thus be implemented in a distributed fashion to obtain the arguments $\{\mathbf{x}_j(n)\}_{j=1}^J$ minimizing the augmented Lagrangian [cf. (3.37) and (3.41)]. In turn, these optimal $\{\mathbf{x}_j(n)\}_{j=1}^J$ can be used to implement the subgradient ascent iteration in (3.30).

The resulting Algorithm 3 embeds an outer iteration (indexed by n) implementing subgradient descent as per (3.30) and an inner iteration (indexed by m for fixed n) implementing coordinate descent as per (3.38)-(3.39) to minimize the augmented Lagrangian. Indeed, steps 4-8 implement (3.38)-(3.39) with steps 6 and 7 representing the interchange of primal variables between neighbors. Strictly speaking step 9 is only true as $M \to \infty$, but even for finite M it provides a reasonable approximation to (3.37) that can be used to find the subgradients in (3.23) and implement the gradient ascent iteration in steps 10 and 11. Steps 12 and 2 communicate the dual variables and step 3 initializes the coordinate descent (inner) iteration.

Different from the dual decomposition in Algorithm 2 convergence of the primal iterates $\{\mathbf{x}_j(n)\}_{j=1}^J$ to the optimal primal arguments $\{\mathbf{x}_j^*\}_{j=1}^J$ as $n \to \infty$ can be guaranteed for the MOM in Algorithm 3 as we summarize in the following property; e.g., [9, Sec. 3.4.4].

Algorithm 3 Method of multipliers

Require: Packet success probabilities to and from neighbors $\mathbf{R}_{c(j)j}$ and $\mathbf{R}_{jc(j)}$

Ensure: Routing probabilities $\bar{\mathbf{s}}_j$

1: **for** $n = 1$ to ∞ **do** {repeat for the life of the network}
2: Receive multipliers $\lambda_{ij}(n)$ and $\mu_{ji}(n)$ from one hop neighbors $\{U_i : i \in c(j)\}$
3: Initial value for coordinate descent: $\mathbf{x}_j(n,0) = \mathbf{x}_j(n-1)$
4: **for** $m = 1$ to M **do**
5: Coordinate descent iteration for $\mathbf{x}_j(n,m)$: $\mathbf{x}_j(n,m) = \arg\min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{A}_j(n,m)$
6: Transmit $w_j(n,m)$, $t_{ij}(n,m)$, and $u_{ji}(n,m)$ to neighbor U_i. Repeat for all $\{U_i : i \in c(j)\}$.
7: Receive $w_i(n,m)$, $t_{ji}(n,m)$, and $u_{ij}(n,m)$ from one hop neighbors $\{U_i : i \in c(j)\}$
8: **end for**
9: Argument minimizing augmented Lagrangian: $\mathbf{x}_j(n) = \mathbf{x}_j(n,M)$
10: Subgradient ascent iteration for $\boldsymbol{\lambda}_j$: $\boldsymbol{\lambda}_j(n) = \boldsymbol{\lambda}_j(n-1) + c[\bar{\mathbf{s}}'_j(n) - \mathbf{u}_j(n)]$
11: Subgradient ascent iteration for $\boldsymbol{\mu}_j$: $\boldsymbol{\mu}_j(n) = \boldsymbol{\mu}_j(n-1) + c[\mathbf{v}_j(n) - w_j(n)\mathbf{1}]$
12: Transmit multipliers $\lambda_{ji}(n+1)$ and $\mu_{ij}(n+1)$ to neighbor U_i. Repeat for all $\{U_i : i \in c(j)\}$.
13: **end for**

Property 2 *Consider implementation of the MOM in Algorithm 3 to solve the optimization problem in (3.12) and let* $\{\mathbf{x}_j^*\}_{j=1}^J$ *denote the arguments minimizing (3.12). Then, for any value of* M, $\lim_{n\to\infty} \mathbf{x}_j(n) = \mathbf{x}_j^*$; *in particular*

$$\lim_{n\to\infty} \bar{\mathbf{s}}_j(n) = \bar{\mathbf{s}}_j^*. \tag{3.42}$$

Property 2 guarantees that the optimal routing probabilities can be obtained by running Algorithm 3. It also establishes that the convergence in (4.16) holds for any number of inner iterations M. A particularly interesting algorithm is obtained by making $M = 1$ leading to the so-called alternating direction MOM. For this algorithm we define the local augmented Lagrangian at time n as

$$\mathcal{A}_j(\mathbf{x}_j, n) = \mathcal{L}_j(\mathbf{x}_j, n) + \frac{c}{2}\left[\|\bar{\mathbf{s}}'_j(n) - \mathbf{u}_j\|^2 + 2\|\mathbf{v}_j(n) - w_j\mathbf{1})\|\|^2 + \|\bar{\mathbf{s}}_j - \mathbf{u}'_j(n)\|^2\right] \tag{3.43}$$

Algorithm 4 Alternating direction method of multipliers

Require: Packet success probabilities to and from neighbors $\mathbf{R}_{c(j)j}$ and $\mathbf{R}_{jc(j)}$

Ensure: Routing probabilities $\bar{\mathbf{s}}_j$

1: **for** $n = 1$ to ∞ **do** {repeat for the life of the network}
2: Receive multipliers $\lambda_{ji}(n)$ and $\mu_{ji}(n)$ from one hop neighbors $\{U_i : i \in c(j)\}$
3: Coordinate descent iteration for $\mathbf{x}_j(n)$: $\mathbf{x}_j(n) = \arg\min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{A}_j(n)$
4: Transmit $w_j(n)$, $t_{ij}(n)$, and $u_{ji}(n)$ to neighbor U_i. Repeat for all $\{U_i : i \in c(j)\}$.
5: Receive $w_i(n)$, $t_{ji}(n)$, and $u_{ij}(n)$ from one hop neighbors $\{U_i : i \in c(j)\}$
6: Subgradient ascent iteration for $\boldsymbol{\lambda}_j$: $\boldsymbol{\lambda}_j(n+1) = \boldsymbol{\lambda}_j(n) + c[\bar{\mathbf{s}}'_j(n) - \mathbf{u}_j(n)]$
7: Subgradient ascent iteration for $\boldsymbol{\mu}_j$: $\boldsymbol{\mu}_j(n+1) = \boldsymbol{\mu}_j(n) + c[\mathbf{v}_j(n) - w_j(n)\mathbf{1}]$
8: Transmit multipliers $\lambda_{ij}(n+1)$ and $\mu_{ij}(n+1)$ to neighbor U_i. Repeat for all $\{U_i : i \in c(j)\}$.
9: **end for**

and define the iteration of the primal variables as

$$\mathbf{x}_j(n+1) = \arg \min_{\mathbf{x}_j \in \mathcal{X}_j} \mathcal{A}_j(\mathbf{x}_j, n) \tag{3.44}$$

with the iteration of the dual variables (multipliers) given as in (3.30). Due to Property 2 the iteration (3.43)-(3.44) and the corresponding Algorithm 4 leads, as $n \to \infty$, to the optimal routing probabilities $\bar{\mathbf{s}}_j^*$.

Remark 7 The algorithms considered in this section assume that packets interchanged for computing the routing matrix $\mathbf{T}$ are received error-free. While it is certainly possible to protect the critical routing information (with, e.g., a powerful error control code) so that this is approximately true, it is not fully with the problem setup consider in this chapter, wherein packets are correctly decoded according to the probabilities in $\mathbf{R}$. An alternative assumption is to assume that routing variables sent from U_j are correctly received by U_i with probability R_{ij} as would be the case if they were included in packet headers. This falls beyond the scope of the present chapter, but it is worth mentioning that there exist asynchronous distributed optimization results that may be used to prove convergence of 2-4 even in this case, under certain conditions [9, Ch. 6].

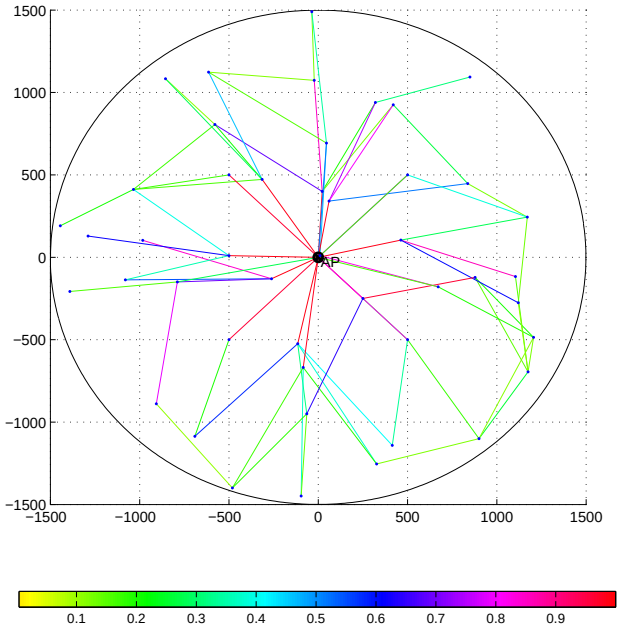

Figure 3.1: Optimal routes for the max-min criterion.

Remark 8 In Algorithms 2-4 node U_j solves a convex optimization problem to minimize the (augmented) Lagrangian at each iteration [cf. (3.29), (3.39), and (3.44)]. The number of variables in these problems is the sum of the dimensions of w_j, $\bar{\mathbf{s}}_j$, and $\mathbf{u}_j$ which amounts to $1 + 2\#[c(j)]$ – with $\#[c(j)]$ denoting the cardinality of $c(j)$. If, as expected, the number of neighbors $\#[c(j)]$ is small, the minimizations in (3.29), (3.39), or (3.44) entail a small computational burden.

3.4 Simulations

Simulations in this section are for a network with $J = 40$ nodes randomly placed on a disc of radius 1.5 km, at whose center is the common access point U_{J+1}. The elements of the packet success probability matrix $\mathbf{R}$ are chosen according to the empirical distribution in [3]. The network considered in the coming experiments is represented in Fig. 3.3. The optimality criterion is max-min rate with corresponding optimal routes given as in Fig. 3.1. The algorithm ran by individual nodes is the alternating direction method of multipliers

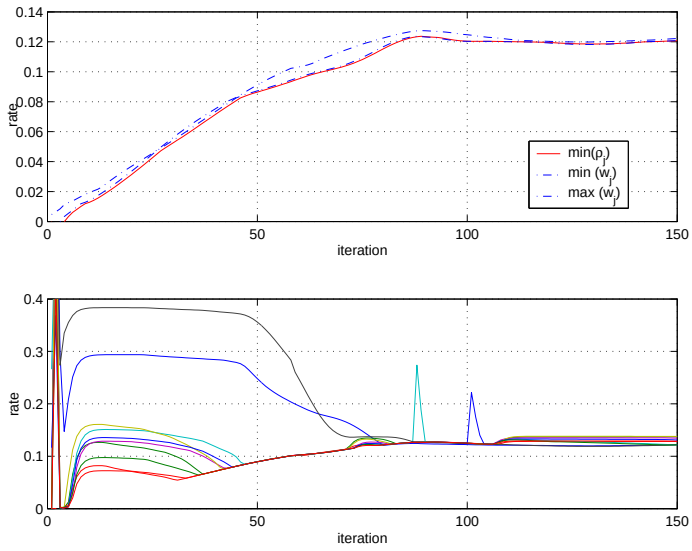

Figure 3.2: Convergence of Algorithm 1 to the max-min optimal routes in Fig. 3.1. After 70 iterations the rate of the most compromised user is within 90% of the optimal rate.

outlined in Algorithm 4.

Alternating direction MOM. We will neglect transmission errors and assume that communication of routing variables between neighbors is always successful. The results of this experiment are summarized in Figs. 3.2-3.5. In Fig. 3.2-(top) we show the smallest and largest value of the local variables w_j. As expected, these variables quickly approach each other due to the constraints $\mathbf{v}_j = w_j \mathbf{1}$; and the minimum rate $\rho_j := \mathbf{r}_j^T \bar{\mathbf{s}}_j - \mathbf{s}_j^T \bar{\mathbf{s}}'_j$ is also closely approximated by the local variables w_j. The rate of convergence is reasonable, since after $n = 150$ iterations the distributed algorithm has converged to the optimal value. Furthermore, we can see that after 70 iterations the rate of the most compromised user is within 90% of the optimal rate. In practice, this last number can be regarded as the time required for convergence. We also plot in Fig. 3.2-(bottom) the path followed by the rate of 10 different representative users with similar conclusions.

Two more interesting experiments shown in Figs. 3.4 and 3.5 illustrate the effect of "topological" changes in the network. In Fig. 3.4 we consider the effect of removing a user from the network. The effect after the first iteration is that the rate of the worst user

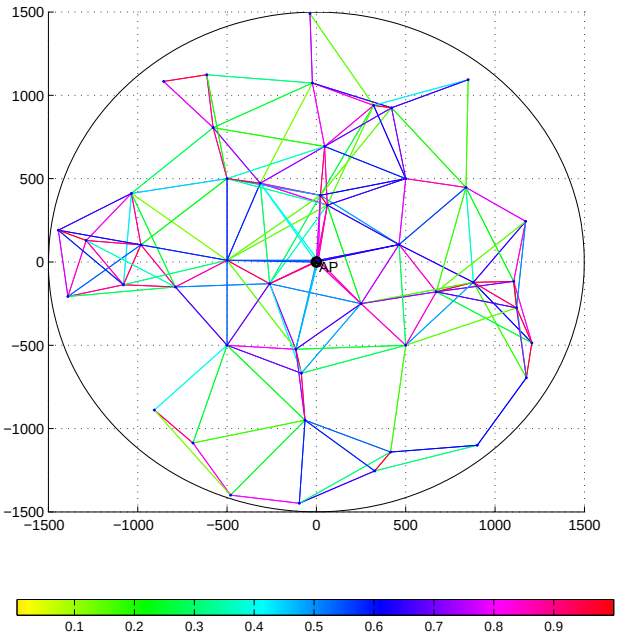

Figure 3.3: Connectivity graph for a network with 40 nodes. The color index represents the value of R_{ij} that is generated according to the empirical distribution in [3].

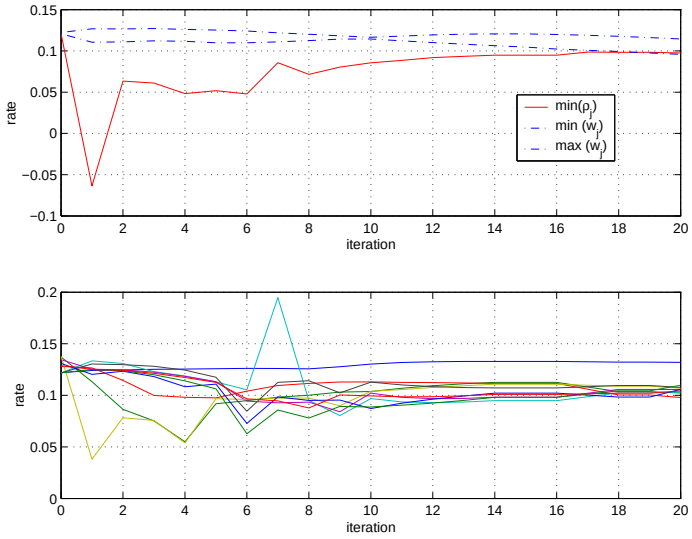

Figure 3.4: Effect of removing a user from the network.

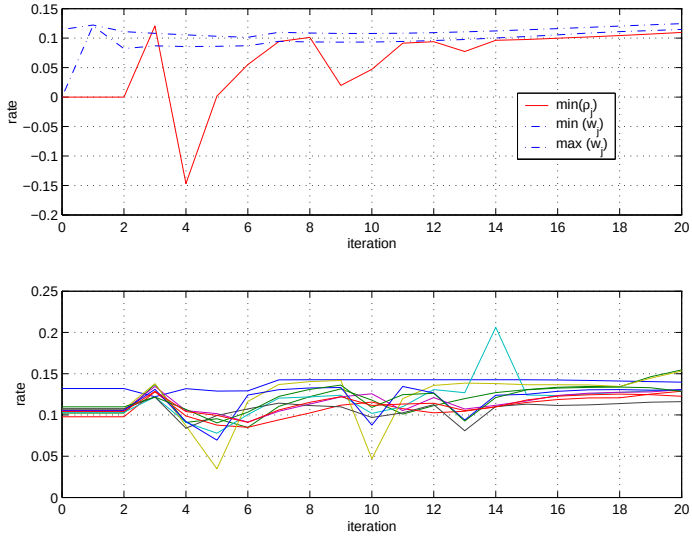

Figure 3.5: Effect of adding a user to the network.

becomes negative corresponding to packets dropped in the routes that were passing through the node removed from the network. The algorithm quickly reroutes the packets leading to positive rates ρ_j in the second iteration, in about 12 iterations Algorithm 2 is again within 90% of optimality and converges to the new optimal throughput after 20 iterations. Similar behavior is observed when we add a new user to the network as shown in Fig. 3.5. We again have a negative rate ρ_j when the new user attempts to send traffic through a congested route. After 8 iterations Algorithm 2 has already found near-optimal routes and converges to a new stable point in about 20 iterations.

Effect of communication errors. Taking into account communication errors as described in Remark 7 leads to Fig. 3.6. Except for the fact that convergence to the optimal solution is slower, taking in the order of 200 iterations for convergence and 90 to reach 90% of the optimal value, the behavior is as with perfect communication links. Omitted to avoid repetition are the experiments showing the effect of adding and dropping terminals in the presence of communication errors. The corresponding simulations are similar to Figs. 3.4 and 3.5. The corresponding number of iterations to attain 90% of optimality are 17 for dropping a user node and 10 for adding a new one.

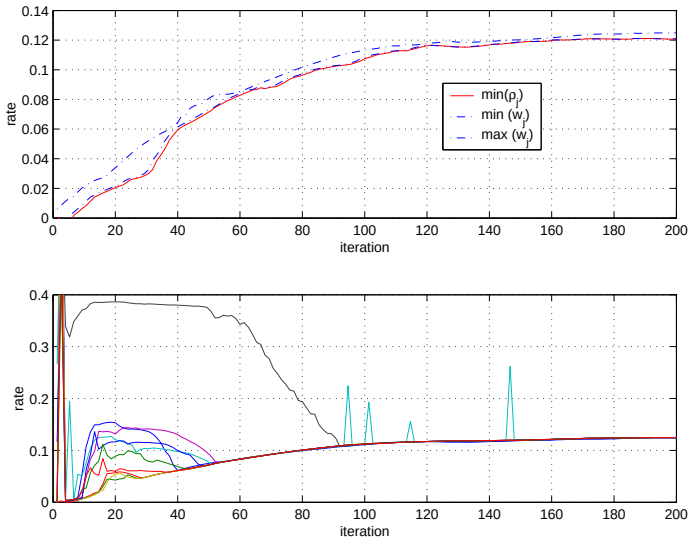

Figure 3.6: Effect of communication errors.

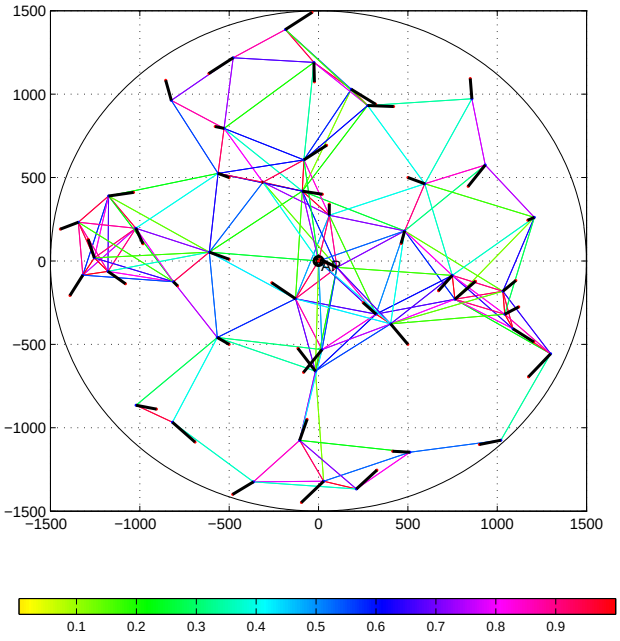

Figure 3.7: Users move 150 meters at random.

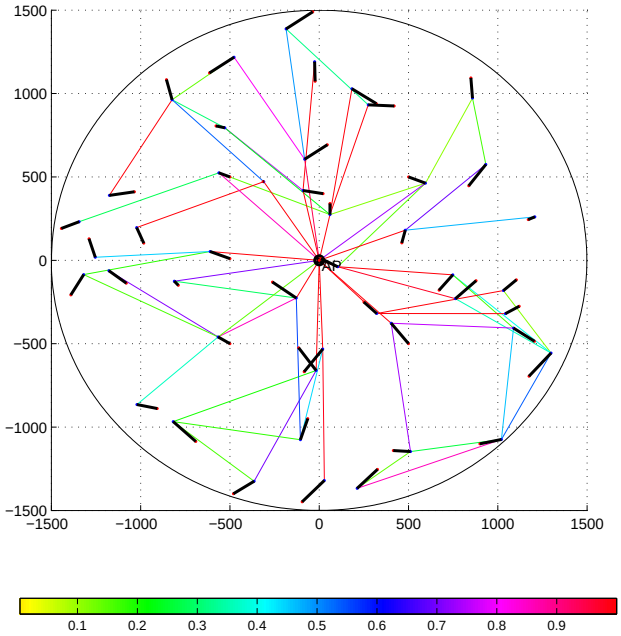

Figure 3.8: Max-min optimal routes for the network in 3.7.

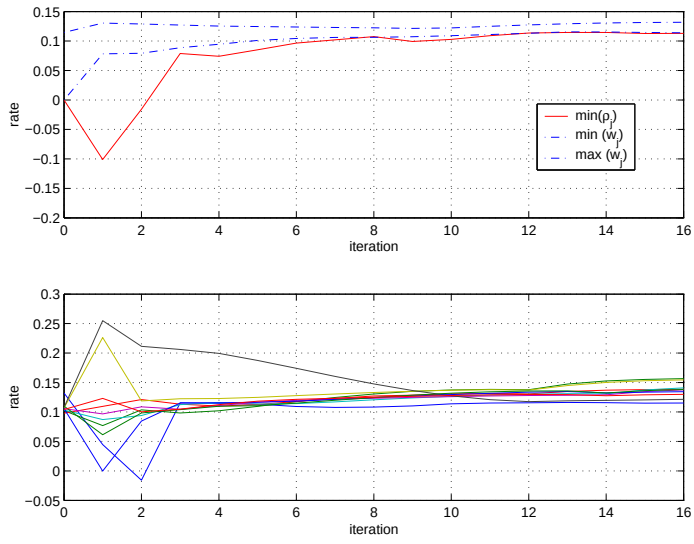

Figure 3.9: Response of Algorithm 1 to user mobility.

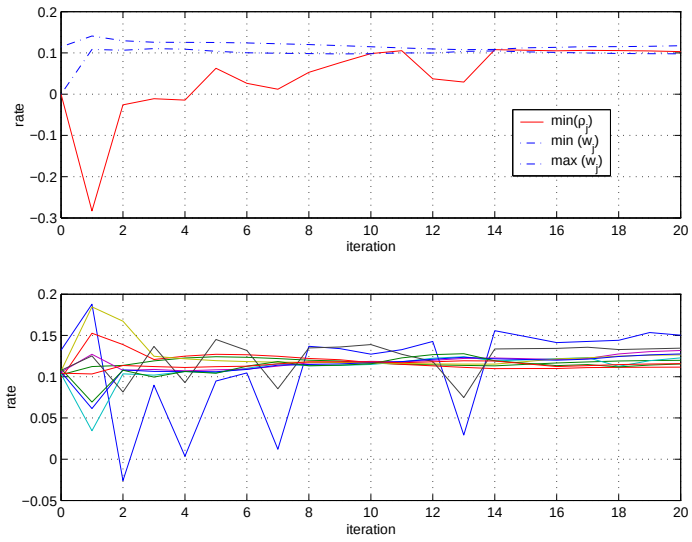

Figure 3.10: Response of Algorithm 1 to user mobility in the presence of communication errors.

Mobility. Among the main motivating reasons behind a distributed implementation is adaptability to a mobile environment. To illustrate this we modify the network in Fig. 3.3 by letting each node move at random with uniform distribution in a square with 300 meter side centered at the original position. This leads to the network in Fig. 3.7. The effect of mobility can be simulated by running Algorithm 2 to find the optimal routes for the network in Fig. 3.7 using the optimal routes in Fig. 3.1 as a initial condition. The results are depicted in Fig. 3.9 for perfect communication of routing variables and in Fig. 3.10 when accounting for communication errors.

In both cases, convergence to the new optimal routes is surprisingly fast taking approximately 8 iterations when communication of routing variables is error-free and 14 when we account for possible communication errors. Intuitively, this happens because optimal routes are robust with respect to modest topology changes.

3.5 Summary

Multi-hop routing in wireless networks holds great promise to improve performance of wireless networks. Building on the results in Chapter 2 that formulate routing problems as convex optimization problems based on the pairwise error probability matrix $\mathbf{R}$, this chapter developed distributed routing algorithms to find rate-optimal routes. Since routing algorithms developed in Chapter 2 cannot be implemented in a distributed fashion we introduced equivalent problems amenable to distributed implementations. Many problems can be cast in the latter formulation including max-min rate, sum-rate, maximum product-rate, and rate-optimal relay networks. In all of these problems additional convex constraints, e.g., minimum acceptable rate or cooperation limit, can be easily incorporated to our framework.

Distributed routing algorithms were obtained via dual decomposition, leading to an iterative algorithm based on communication with one-hop neighbors only. Since in many cases of interest dual decomposition iterates do not necessarily converge to the optimal routing matrix, we adopted two well-known regularization approaches, namely the method of multipliers (MoM) and the alternating direction MoM. Convergence of the MoM and the alternating direction MoM algorithms to the optimal routing matrix is guaranteed under mild conditions. Of particular practical importance is the guaranteed convergence in the presence of communication errors.

Simulations corroborated that the MoM is a robust algorithm quickly converging to the optimal routes. We further showed that the resulting algorithms are fast to respond to addition and removal of terminals as well as to changes in the pairwise error probability matrix brought in by, e.g., node mobility.

Chapter 4

Cooperative diversity in multiple access channels

Rich scattering of electromagnetic waves propagating through physical environments generates complex interference patterns. As such patterns go through maxima and minima, large variations in energy adversely affect wireless reception and thus deteriorate error probability performance of wireless communication systems. By providing multiple channels with independent (or at least uncorrelated) variations in time, frequency and/or space, *diversity* techniques offer well-appreciated countermeasures mitigating such (so called fading) effects. With the deployment of multiple antennas effecting space diversity we create copies of the transmitted signal either at the receiver, at the transmitter or both. In time or frequency (a.k.a. Doppler or multipath) diversity systems, we exploit the natural property of wireless channels to vary over time or frequency. The benefits of diversity are significant. In a typical (wireline) additive white Gaussian noise (AWGN) channel the error probability decays exponentially as the received signal-to-noise-ratio (SNR) increases; i.e., error effects decrease as $e^{-\mathrm{SNR}}$. A wireless Rayleigh fading channel however, exhibits errors decaying as SNR^{-1}. A κ^{th}-order diversity channel entails κ uncorrelated channels and exhibits error probability which decreases as $\mathrm{SNR}^{-\kappa}$. A pertinent definition when analyzing diversity

enabling protocols is the diversity order.

$$\eta := \lim_{\gamma \to \infty} \frac{\log[P_e(\gamma)]}{\log(\gamma)}. \tag{4.1}$$

Needless to say the gap between the exponential decay in wireline channels and the inversely linear decay in wireless channel is enormous. Considering that for sufficiently large κ the $\mathrm{SNR}^{-\kappa}$ and $e^{-\mathrm{SNR}}$ functions are not very different, the value of diversity is clear: it can close the error performance gap between wireline and wireless channels.

Spatial and time-frequency diversity systems are at opposite ends of a deployment cost versus reliability curve. Spatial diversity is reliable but comes with hardware cost. Time-frequency diversity on the other hand exploits natural phenomena that may or may not be present in a particular link and is thus less reliable even if it comes for free when available. User *cooperation* is an alternative form of diversity which aims to strike a balance in this curve by providing diversity more reliable than natural time-frequency variations yet without requiring deployment of additional antennas. The basic idea is to have single-antenna terminals share information and cooperate in relaying it to intended destinations. If properly designed, cooperative protocols involving κ terminals can achieve κ^{th}-order diversity relying on relatively inexpensive software modifications of existing wireless protocols.

Since its introduction [51, 96, 97], researchers in signal processing, wireless communications and information theory have contributed major advancements to explore and realize the potential of cooperative networks. The contribution of this thesis in this area is in understanding user cooperation for multiple access (MA) over fixed as well as random access (RA) channels.

Notation: The canonical basis of $\mathbb{R}^N$ will be denoted as $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ so that the $N \times N$ identity matrix can be written as $\mathbf{I}_N := [\mathbf{e}_1, \ldots, \mathbf{e}_N]$. The all-one and all-zero vectors in $\mathbb{R}^N$ will be denoted as $\mathbf{1}_N := [1, \ldots, 1]^T$ and $\mathbf{0}_N := [0, \ldots, 0]^T$, respectively.

4.1 Single source cooperation (SSC)

The core idea behind user cooperation is to create a virtual antenna array (VAA) for transmission by means of data sharing between users. With reference to Fig. 4.1, consider

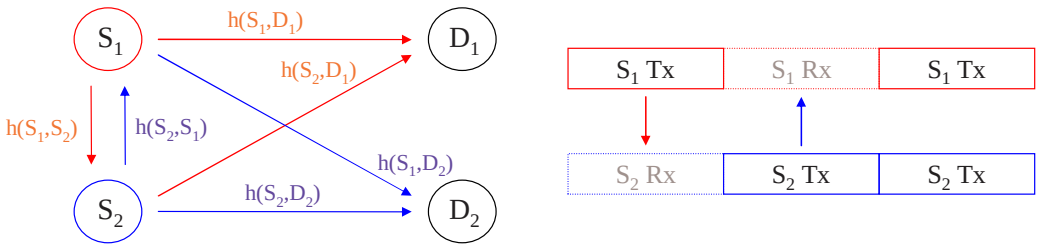

Figure 4.1: Source terminals S_1 and S_2 cooperate in transmitting to their respective destinations D_1 and D_2 by creating a distributed virtual antenna array (VAA).

source S_1 (S_2) sending a data packet $\mathbf{d}_1$ ($\mathbf{d}_2$) to destination D_1 (D_2) through the wireless Rayleigh flat fading channel $h(S_1, D_1)$ [$h(S_2, D_2)$]. Due to the broadcast nature of the wireless channel, $\mathbf{d}_1$ transmitted by S_1 is not only received by D_1 but also by S_2 and D_2 through corresponding channels $h(S_1, S_2)$ and $h(S_1, D_2)$. Thus, if we let S_2 repeat the signal received from S_1 and vice versa, both destinations receive two independent copies of $\mathbf{d}_1$ and $\mathbf{d}_2$. Forgetting for a moment the channel $h(S_1, S_2)$ between sources, D_1 receives data from a 2×1 multiple input single output (MISO) channel which is capable of providing second-order diversity [4].

Even though similar, there are important differences between VAAs and MISO systems with multiple co-located antennas. One difference is that wireless terminals are half-duplex, and as such they cannot transmit and receive over the same frequency at the same time. This practical limitation is rooted in the need to isolate transmitter and receiver in order to avoid feedback from the transmitter to the receiver radio-frequency (RF) front end. If the terminal size is not enough to provide spatial isolation, this has to be achieved in time and/or frequency. The implication is that cooperation protocols have to follow a scheme like the one depicted in Fig. 4.1 in which we have a slot assigned to S_1's transmission, a second slot assigned to S_2's and a third slot for the cooperative transmission of the other terminal's data. Comparing this scheme with space-time codes [4] we recognize that different from MISO channels the diversity advantage of VAAs comes at the price of bandwidth increase. It is worth noting that this does not necessarily imply a penalty in communication rate, because the decrease in the amount of forward error correction (FEC) and/or number of

re-transmissions required can compensate for the bandwidth expansion [96].

A second difference is that in VAAs we cannot ignore the channel $h(S_1, S_2)$ between sources. To appreciate its effects, let $\hat{\mathbf{d}}_1$ denote S_2's estimate of $\mathbf{d}_1$ and consider the signals received by the destination D_1:

$$\begin{aligned} \mathbf{y}_{11} &= \sqrt{P} h(S_1, D_1)\mathbf{d}_1 + \mathbf{w}_{11}, \\ \mathbf{y}_{12} &= \sqrt{P} h(S_2, D_1)\hat{\mathbf{d}}_1 + \mathbf{w}_{12}, \end{aligned} \tag{4.2}$$

where $\mathbf{w}_{11}$ and $\mathbf{w}_{12}$ denote AWGN terms and P is the transmitted power. It is a surprising result that if D_1 uses a maximum ratio combiner (MRC) for estimating $\mathbf{d}_1$ as (* stands for conjugation and $\|x\|$ for the magnitude of x)

$$\hat{\mathbf{d}}_1^{\mathrm{MRC}} = \arg\min_{\mathbf{d}_1} \left\| h^*(S_1, D_1)\mathbf{y}_{11} + h^*(S_2, D_1)\mathbf{y}_{12} - \sqrt{P}\left[|h(S_1, D_1)|^2 + |h(S_2, D_1)|^2\right]\mathbf{d}_1 \right\|, \tag{4.3}$$

then the diversity order of this two-branch VAA is only one. The reason for the lack of diversity in this so called decode and forward (DF) strategy is that the VAA error probability is dominated by the error probability in the link $S_1 \to S_2$.

While DF does not achieve diversity, three alternative strategies do achieve this goal:

[S1] *Selective forwarding (SF):* Instead of always repeating $\mathbf{d}_1$, S_2 will repeat the packet only if it is successfully decoded i.e., if $\hat{\mathbf{d}}_1 = \mathbf{d}_1$. This strategy is more complex than DF because it requires FEC decoding followed by a cyclic redundancy code (CRC) check to detect possible errors at S_2.

[S2] *Amplify and forward (AF):* A seemingly simple alternative is to let S_2 amplify the analog-amplitude signal received from S_1. That is, the signal $\mathbf{y}_{21} = h(S_1, S_2)\mathbf{d}_1 + \mathbf{w}_{21}$ received by S_2 is transmitted after amplification as $A\mathbf{y}_{21}$. The amplification factor satisfies

$$A^2 = \frac{P}{P|h(S_1, S_2)|^2 + N_0}, \tag{4.4}$$

so that the power of the signal transmitted by S_2 is equal to P.

[S3] *Cooperative (C) MRC:* While the strategies [S1] and [S2] require operations at the cooperating terminal, a different approach is to adopt DF at the cooperating terminal

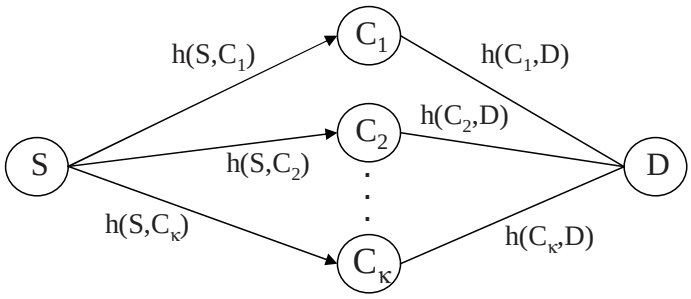

Figure 4.2: Multi-branch cooperation.

but use a weighted version of the MRC demodulator in (4.3)

$$\hat{\mathbf{d}}_1^{\text{CMRC}} = \arg\min_{\mathbf{d}_1} \left\| \alpha_{11}\mathbf{y}_{11} + \alpha_{12}\mathbf{y}_{12} - \sqrt{P}\left[\alpha_{11}h(S_1, D_1) + \alpha_{12}h(S_2, D_1)\right]\mathbf{d}_1 \right\|. \quad (4.5)$$

By properly selecting α_{11} and α_{12} as functions of $h(S_1, D_2)$, $h(S_2, D_2)$ and $h(S_1, S_2)$ the so called C-MRC in (4.5) can be shown to achieve second-order diversity [15].

Each of the strategies [S1]-[S3] has its own merits. SF is the simplest one from the perspective of the destination but strains the digital processor at the cooperating terminal; also, even if the packet is not correctly decoded there is still some information about $\mathbf{d}_1$ in the signal received at the cooperator that is not conveyed to the destination. When the link between sources ($S_1 \to S_2$) is expected to be much better than the links between sources and destination ($S_1, S_2 \to D_1$), S_2 will almost always correctly decode $\mathbf{d}_1$ making SF the method of choice for this case. AF requires minimal processing at the cooperating terminal, but necessitates storage of the analog-amplitude received signal thus straining memory resources. AF is appealing when the cooperating terminal is located close to the destination so that the link from the cooperating terminal to the destination ($S_2 \to D_1$) is strong and the link $S_1 \to S_2$ is comparable to the link $S_1 \to D_1$. Use of C-MRC for decoding DF relayed signals is the simplest strategy from the perspective of the cooperating terminal. Its drawback is that the channel realization $h(S_1, S_2)$ has to be transmitted to the destination since it is needed to compute α_{11} and α_{12}. If this can be accomplished by transmitting a few bits the overhead is not significant.

Pairwise cooperation can be generalized to groups of terminals. For a group of κ co-

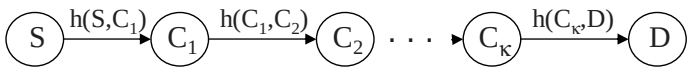

Figure 4.3: Multi-hop cooperation.

operating terminals we can build a protocol using any of the strategies [S1]-[S3] to achieve κ^{th}-order diversity. This may not be always the best approach considering that in cooperative networks – sometimes also referred to as relay networks – there is a tradeoff between multi-branching (see Fig. 4.2) and multi-hopping (see Fig. 4.3). In multi-hopping, the source packet is relayed through a cascade of cooperating terminals; while not providing diversity, this approach saves energy by exploiting the smaller pathloss between cooperators as compared to the pathloss from source to destination. In multi-branching, the packet is relayed to κ cooperators that retransmit the packet to the destination; this provides diversity but does not benefit from pathloss reduction. The configuration offering desirable tradeoffs in a general network is a combination of multi-hop and multi-branch cooperation [75].

Remark 9 We have introduced only simple concepts of SSC necessary to study cooperation in multiple fixed and random access channels. Among topics we did not cover due to space limitations is the aforementioned bandwidth penalty VAAs incur relative to MISO systems with co-located antennas. A possible remedy is resorting to (an e.g., turbo) coded cooperation whereby the source transmits the first sub-block of the code, the cooperating terminal decodes the signal using only this first sub-block and, if successful, transmits the second sub-clock of the turbo code. This does not incur bandwidth expansion to implement cooperation but requires coding at the relays which expands bandwidth, even though the latter is arguably needed anyways [40]. An additional issue is the use of coherent versus non-coherent reception. The use of non-coherent modulation in cooperative networks and its diversity benefits are reported in [17] and [118]. Fundamental performance limits of cooperative links are closely related to the capacity of the relay channel, the evaluation of which remains an open problem in information theory [20]. It has been shown that the bandwidth penalty of cooperative protocols is not inherent to the relay channel but is due to the use of repetition coding [5]. In the low-power regime, achievable rates and optimum

resource allocation issues for the relay channel have been studied in [14] and [126]. For the Gaussian relay channel it is also known that DF and AF relay strategies can be outperformed by a quantize and forward (QF) scheme, whereby the cooperating relay forwards a quantized version of the source signal [47].

4.2 Cooperation in multiple access channels

In a multiple access channel, "good performance" is quantified by low error rate, high spectral efficiency and low complexity. On the other hand, multiplicative fading induced by the propagation environment and additive noise effects at the receiving end, render it impossible to optimize one metric without sacrificing the others. A universal system design should, therefore, be flexible to tradeoff among error performance, spectral efficiency and complexity.

The name Multi-source cooperation (MSC) to denote cooperation between terminals of a multiple access network was introduced in [100] to improve bandwidth efficiency and diversity order. A two-phase MSC system with distributed convolutional coding (DCC) was reported in [117], along with a simple design of interleavers to maximize the diversity order of simple error events. A distributed trellis coded modulation (DTCM) based MSC system approaching the bandwidth efficiency of a non-cooperative time division multiple access (TDMA) system was developed in [116]. Assuming slow block fading Rayleigh channels with binary transmission, the maximum achievable diversity order effected by error control coding (ECC) in MSC networks with K users is $\eta = \min(d_{\min}, \lfloor 1 + K(1 - R_c) \rfloor)$, where $d_{\min}$ and R_c denote respectively the minimum (free) distance and the ECC rate [116, 117].

However, viewing K transmitting users as a virtual antenna array suggests that the attainable diversity order could be as high as K. Clearly, if $R_c > 1/K$, then MSC with ECC cannot achieve this maximum diversity order. This is actually inherent to the diversity properties of ECC. On the other hand, complex field coding (CFC) applied to co-located (N_t, N_r) multi-antenna systems is known to achieve transmission rate of N_t symbols per channel use with diversity order as high as the product of the number of transmit-receive antennas, i.e., $\eta_{\max} = N_t N_r$ [60, 125]. This motivates adoption of distributed CFC (DCFC)

in MSC networks to effect diversity order equal to the number of cooperating users K.

This chapter introduces a general MSC framework with full-diversity, flexible spectral efficiency and controllable decoding complexity. Users are grouped in clusters that are separated with code division multiple access (CDMA). Within each cluster users cooperate to reach the access point (AP), implementing MSC according to a two-phase TDMA protocol (Section 4.3). Our first contribution is to show that the diversity order of MSC over fading channels coincides with the diversity order when the links between cooperating users are error-free (Section 4.3.1). As the latter can be thought as a single user transmission over multiple input - single output (MISO) channels, two implications of this result are: i) the diversity order of repetition coding is $\eta_{\text{RC}} = 2$; and ii) the maximum diversity order of distributed ECC is $\eta_{\text{DECC}} = \min(d_{\min}, \lfloor 1+K(1-R_c) \rfloor)$. The second contribution is to establish that DCFC-based MSC enables diversity order equal to the number of users, $\eta_{\text{DCFC}} = K$ (Section 4.3.2). We further address cluster separation and demonstrate that when the number of clusters is larger than the spreading gain, flexible MSC protocols emerge trading off spectral efficiency, error performance and complexity (Section 4.4). While coding gain is affected in this under-spread case the diversity order is not, thus enabling MSC protocols to achieve full diversity at maximum spectral-efficiency equal to that of non-cooperative networks. By adjusting the number of cooperating users, MSC encoder, spreading gain and/or the number of clusters, our general MSC framework is flexible to tradeoff among spectral-efficiency, decoding complexity and diversity.

4.3 Multi-source cooperation

Consider the cooperative multiple access (MA) setup of Fig. 4.4 in which the set of active users is divided into L clusters $\{\mathcal{U}_l\}_{l=1}^{L}$. In each cluster, K_l users $\{U_{lk}\}_{k=1}^{K_l}$ cooperate in transmitting symbol blocks $\mathbf{s}_{lk} := [s_{lk1}, \ldots, s_{lkN}]^T$ of size $N \times 1$ to the AP that we write as U_{00}. We assume that $\mathbf{s}_{lk}$ contains a cyclic redundancy check (CRC) code allowing detection of correctly received packets. We let $\mathbf{h}_{l_1k_1,l_2k_2} := h_{l_1k_1,l_2k_2}\mathbf{1}_N$ denote the block Rayleigh fading channel between users $U_{l_1k_1}$ and $U_{l_2k_2}$; and $\mathbf{h}_{lk} := h_{lk}\mathbf{1}_N$ the one between U_{lk} and the AP. We further assume that these channels are uncorrelated and adopt the convention

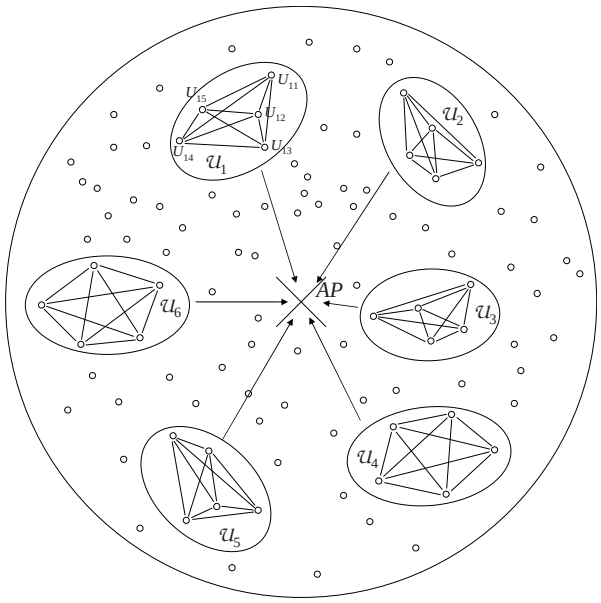

Figure 4.4: Multiple access (MA) channel is divided in cooperating clusters.

$\mathbf{h}_{lk,lk} \equiv \mathbf{1}_N$. Let us postpone to Section 4.4 the issue of cluster separation and focus on the operation of a single cluster, for which we set $L = 1$ and drop the cluster subscript l to simplify notation.

Supposing that frame synchronization has been established, TDMA is used to separate users per cluster as depicted in Fig. 4.5. The MSC protocol consists of two phases each taking place over K slots. With symbol duration T_s, unit-energy pulse waveform $p(t)$ with non-zero support T_s and amplitude A, the waveform transmitted by source U_k during phase-1 is

$$x_k^{(1)}(t) = A \sum_{n=1}^{N} s_{kn} p\left[t - ((k-1)N + n)T_s\right], \qquad t \in [0, KNT_s). \tag{4.6}$$

The waveforms $\{x_k^{(1)}(t)\}_{k=1}^K$ propagate through the shared wireless interface so that over a burst of KNT_s seconds each user in the cluster, say U_k, has available the waveform $y_k^{(1)}(t) = \sum_{j=1}^K h_{j,k}(t) x_j^{(1)}(t) + n(t)$, where $n(t)$ denotes AWGN with double-sided spectral density $N_0/2$. The waveform $y_k^{(1)}(t)$ is subsequently match filtered yielding samples

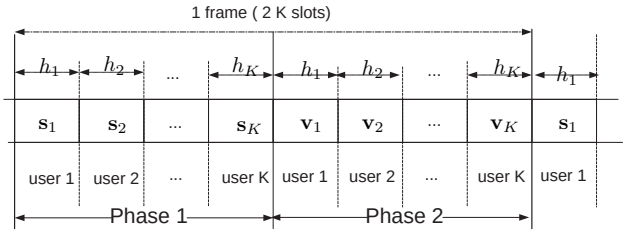

Figure 4.5: TDMA structure of an MSC protocol for a cluster with K active users.

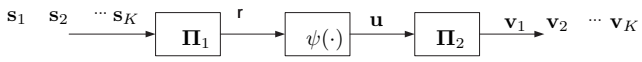

Figure 4.6: Encoder and interleaving modules of each cooperating user.

$y_{k,jn} = \int_0^{T_s} y_k^{(1)}[t-((j-1)N+n)T_s]p^*(t)dt$. Upon defining the aggregate $KN \times 1$ transmitted and received blocks $\mathbf{s} := [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T$ and $\mathbf{y}_k^{(1)} := [y_{k,11}, \ldots, y_{k,1N}, y_{k,21}, \ldots, y_{k,KN}]^T$, the noise vector $\mathbf{n}_k^{(1)} := [n_{k,11}, \ldots, n_{k,KN}]^T$ and the diagonal channel matrix $\mathbf{D}_k^{(1)} := \text{diag}(\mathbf{h}_{k,1}^T, \ldots, \mathbf{h}_{k,K}^T)$, the input-output relationship per user U_k during phase-1 is

$$\mathbf{y}_k^{(1)} = A\mathbf{D}_k^{(1)}\mathbf{s} + \mathbf{n}_k^{(1)}, \qquad k = 0, 1, \ldots, K, \tag{4.7}$$

where by convention $\mathbf{h}_{k,k} \equiv \mathbf{1}_N \ \forall k$, $n_{k,kn} \equiv 0$ for $n \in [1, N]$ and we recall that $\mathbf{y}_0^{(1)}$ corresponds to the received block at the AP. For future use, we note that the transmit SNR is $\gamma := A^2/N_0$ and the average SNR in the $U_k \to U_j$ link is $\gamma_{k,j} := (A^2/N_0)\text{E}[h_{k,j}^2] = \gamma\text{E}[h_{k,j}^2]$.

Notice that by the end of phase-1 every user has available information about the symbol blocks of all users in the cluster $\mathcal{U}$. User U_k ($k > 0$) estimates the joint block $\mathbf{s}$, with entries drawn from a signal constellation $\mathcal{S}$, using the maximum likelihood (ML) decoder [c.f. (4.7)]

$$\hat{\mathbf{s}}_k = \arg\min_{\mathbf{s}\in\mathcal{S}^N} \|\mathbf{y}_k^{(1)} - A\mathbf{D}_k^{(1)}\mathbf{s}\|. \tag{4.8}$$

If symbols in $\mathbf{s}$ are uncoded, then (4.8) amounts to symbol-by-symbol detection since $\mathbf{D}_k^{(1)}$ is diagonal; whereas if the individual blocks $\mathbf{s}_k$ are protected with ECC, then (4.8) implements block detection. Since not all users in $\mathcal{U}$ decode $\mathbf{s}$ correctly, we define the set of those that

do as

$$\mathcal{D} := \{U_k \mid \hat{\mathbf{s}}_k = \mathbf{s}\} \subseteq \mathcal{U}. \tag{4.9}$$

Users in $\mathcal{D}$ proceed to phase-2, but before transmission they process $\mathbf{s}$ as shown in Fig. 4.6. The aggregate block $\mathbf{s}$ is fed to an interleaver, Π_1, yielding the vector $\mathbf{r} = \mathbf{\Pi}_1\mathbf{s}$. The interleaved block $\mathbf{r}$ is then encoded with a function $\psi(\cdot)$ to obtain $\mathbf{u} = \psi(\mathbf{r})$, which is subsequently fed to a second interleaver, Π_2, to obtain the block $\mathbf{v} = \mathbf{\Pi}_2\mathbf{u}$. This processing per user in $\mathcal{D}$ can be summarized as

$$\mathbf{v} = \mathbf{\Pi}_2\mathbf{u} = \mathbf{\Pi}_2\psi(\mathbf{r}) = \mathbf{\Pi}_2\psi(\mathbf{\Pi}_1\mathbf{s}). \tag{4.10}$$

Since all operations in (4.10) preserve dimensionality, we have that the blocks $\mathbf{v}, \mathbf{u}, \mathbf{r} \in \mathbb{R}^{NK\times 1}$, the matrices $\mathbf{\Pi}_1, \mathbf{\Pi}_2 \in \mathbb{R}^{NK\times NK}$ and the encoder $\psi : \mathbb{R}^{NK\times 1} \to \mathbb{R}^{NK\times 1}$.

Each user in $\mathcal{D}$ transmits again in a TDMA fashion an $N \times 1$ sub-block of the block $\mathbf{v} := [v_{11}, \ldots, v_{1N}, v_{21}, \ldots, v_{KN}]^T$. Specifically, U_k transmits the sub-block $\mathbf{v}_k := [v_{k1}, \ldots, v_{kN}]^T$ using the waveform

$$x_k^{(2)}(t) = A\sum_{n=1}^{N} v_{kn}\, p\left[t - ((k-1)N + n)T_s\right], \qquad t \in [0, KNT_s). \tag{4.11}$$

The AP receives $x_k^{(2)}(t)$ from all users $U_k \in \mathcal{D}$ and nothing from the remaining users $U_k \notin \mathcal{D}$. To describe this reception, define the $N \times 1$ channel vector $\tilde{\mathbf{h}}_k := \mathbf{h}_k$, if $U_k \in \mathcal{D}$; and $\tilde{\mathbf{h}}_k := \mathbf{0}_N$, otherwise. Using this model, the block of samples at the matched filter output of the AP in phase-2 is given by

$$\mathbf{y}_0^{(2)} = A\tilde{\mathbf{D}}\mathbf{\Pi}_2\psi(\mathbf{\Pi}_1\mathbf{s}) + \mathbf{n}_0^{(2)}, \tag{4.12}$$

where $\tilde{\mathbf{D}} := \mathrm{diag}(\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_K)$. The blocks $\mathbf{y}_0^{(1)}$ and $\mathbf{y}_0^{(2)}$ received in the two phases can be combined in the aggregate input-output relationship

$$\begin{bmatrix} \mathbf{y}_0^{(1)} \\ \mathbf{y}_0^{(2)} \end{bmatrix}_{2KN\times 1} = A\begin{bmatrix} \mathbf{D}_0^{(1)}\mathbf{s} \\ \tilde{\mathbf{D}}\mathbf{\Pi}_2\psi(\mathbf{\Pi}_1\mathbf{s}) \end{bmatrix} + \begin{bmatrix} \mathbf{n}_0^{(1)} \\ \mathbf{n}_0^{(2)} \end{bmatrix} \tag{4.13}$$

that the AP relies on to jointly decode $\mathbf{s}$. Note that all channels are assumed invariant over the duration of the two phases. Furthermore, since we transmit KN symbols in $2KN$ time slots, the spectral efficiency of single-cluster MSC is $\xi = 1/2$.

If all the diagonal entries of $\mathbf{D}_0^{(1)}$ and $\tilde{\mathbf{D}}$ were Rayleigh distributed, then (4.13) could be thought as the input-output relationship of a coded MISO system transmitting $\mathbf{s}$ and its encoded version over K antennas with only one antenna transmitting N symbol periods in a cyclic fashion. Equivalently, (4.13) could model an $1 \times K$ single input - multiple output (SIMO) channel with K receive antennas, or, a single antenna time-selective block fading channel with K degrees of freedom. In any event, the diversity order can be evaluated once the encoder $\psi(\cdot)$ has been specified. Having as elements the (Rayleigh) block fading channels $\mathbf{h}_k$ between U_k and the AP, $\mathbf{D}_0^{(1)}$ is Rayleigh distributed. However, the second phase equivalent channels $\tilde{\mathbf{h}}_k$ have a distribution that depends on the probability of successful decoding – recall that $\tilde{\mathbf{h}}_k := \mathbf{h}_k$, if $U_k \in \mathcal{D}$; and $\tilde{\mathbf{h}}_k := \mathbf{0}_N$, otherwise. While this distribution is not difficult to characterize it is certainly not Rayleigh; hence, $\tilde{\mathbf{D}}$ is not Rayleigh distributed either. We will prove later in Theorem 7 that even if $\tilde{\mathbf{D}}$ is not Rayleigh distributed the diversity order of MSC protocols coincides with the diversity order when $\tilde{\mathbf{D}}$ is Rayleigh distributed.

Defining a particular MSC protocol amounts to specifying the triplet $(\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1)$ in (4.10). The diversity enabled by any MSC protocol is mainly determined by the encoder $\psi(\cdot)$; while the interleavers $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ distribute relayed symbols to different channels in order to effect the diversity order enabled by $\psi(\cdot)$. In this sense, the unifying framework presented in this section subsumes a number of existing cooperative protocols as special cases. Three of them are highlighted next.

[C1] Distributed repetition coding: Setting the permutation matrices $\mathbf{\Pi}_1 = \mathbf{\Pi}_2 = \mathbf{I}$ and selecting the encoder as

$$\psi_R(\mathbf{r}) = \psi_R(\mathbf{s}) = \psi_R([\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T) := [\mathbf{s}_2^T, \ldots, \mathbf{s}_K^T, \mathbf{s}_1^T]^T, \tag{4.14}$$

reduces the input-output relationships (4.7) and (4.12) to those encountered with MSC based on repetition coding whereby U_k repeats U_{k-1}'s frame for $k \neq 1$ and U_1 repeats U_K's frame [96].

[C2] Distributed ECC: Let $\varphi(\cdot)$ be the function mapping a symbol vector over the *Galois field* $GF(m)$, to a channel codeword and $\varphi^{-1}(\cdot)$ the corresponding de-mapping

function. Let also $\mathbf{P}$ denote the generator matrix of a channel encoder $\psi_P(\mathbf{r}) := \varphi(\mathbf{P}\varphi^{-1}(\mathbf{r}))$ with multiplication defined over $GF(m)$. If $\mathbf{P}$ generates a Reed-Solomon code, $\psi_P(\mathbf{r})$ specifies the MSC protocol in [100]; whereas if $\mathbf{P}$ generates a convolutional code, $\psi_P(\mathbf{r})$ gives rise to the DCC based MSC protocol in [117]. To effect the diversity, $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ have to be tailored for each chosen ECC [117]. Other channel codes, including distributed trellis coded modulation [116], are also possible choices.

[C3] **Distributed CFC:** Consider now the encoder $\psi_\Phi(\mathbf{r}) = \mathbf{\Phi r}$, where $\mathbf{\Phi}$ is a block diagonal matrix with complex entries and multiplication is over the *complex field*. This selection of $\psi(\cdot)$ corresponds to distributed complex field coding (DCFC) that we will elaborate on later in this section.

Remark 10 The set $\mathcal{D}$ of users that correctly decoded $\mathbf{s}$ does not need to be known to the cooperating users. This feature is important in practical deployment and can be readily verified by inspecting the encoder steps in (4.10) which clearly do not depend on $\mathcal{D}$.

4.3.1 Diversity Analysis

The diversity order η in (4.1) enabled by the generic MSC protocol we described so far depends on the encoder $\psi(\mathbf{r})$ in (4.10). Even though this precludes assessment of the diversity order without referring to a specific $\psi(\mathbf{r})$, we can obtain a general result by relating the MSC setup with an equivalent *single-user* transmission of $[\mathbf{s}^T, \mathbf{v}^T]^T$ (comprising the systematic and parity symbols) over a single-antenna block fading channel $\mathbf{D}_0^{(1)}$. If the links between cooperating users are error-free, then $\mathcal{D} \equiv \mathcal{U}$ and (4.12) becomes

$$\mathbf{y}_0^{(2)} = \mathbf{D}_0^{(1)}\mathbf{\Pi}_2\psi(\mathbf{\Pi}_1\mathbf{s}) + \mathbf{n}_0^{(2)}. \tag{4.15}$$

The AP can, therefore, decode $\mathbf{s}$ as if it were transmitted by a single user over a single time-selective fading channel. The only difference between (4.12) and (4.15) is that the channel matrix $\tilde{\mathbf{D}}$ in (4.12) is replaced by $\mathbf{D}_0^{(1)}$ in (4.15). From a statistical point of view, the only difference between these two models is the probability distribution of $\mathbf{D}_0^{(1)}$ and $\tilde{\mathbf{D}}$; from a practical perspective, we can think of (4.15) as the limiting case of (4.12) with

perfect decoding in user-to-user links. Regardless of the interpretation, the important point is stated in the following theorem.

Theorem 7 *Let $\eta[\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1]$ be the diversity order of the MSC protocol with input-output relations given by (4.7) and (4.12). Likewise, let $\beta[\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1]$ be the diversity order of the equivalent single user protocol with input-output relations (4.7) and (4.15). Then, for any encoder $\psi(\cdot)$ and permutation matrices $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$, it holds that*

$$\eta[\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1] = \beta[\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1]. \tag{4.16}$$

To prove Theorem 7 we need two lemmas. In Lemma 2 we assess the diversity order conditioned on the set of decoders $\mathcal{D}$ (see Appendix A for the proof). In Lemma 3, we characterize the probability distribution of $\mathcal{D}$ as $\gamma \to \infty$ (see Appendix B for the proof).

Lemma 2 *If $\eta(\mathcal{D}) := \lim_{\gamma\to\infty} \log[P_e(\gamma|\mathcal{D})]/\log(\gamma)$ denotes the diversity order of the MSC protocol in Theorem 7 conditioned on the decoding set $\mathcal{D}$, then*

$$\eta(\mathcal{D}) \geq \max[0; \beta - (K - |\mathcal{D}|)], \tag{4.17}$$

where $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$ and $\beta := \beta[\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1]$.

Lemma 3 *The probability $\Pr(\mathcal{D})$ of the decoding set $\mathcal{D}$ is such that*

$$\lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathcal{D})\right]}{\log(\gamma)} = -\left(K - |\mathcal{D}|\right). \tag{4.18}$$

Lemma 2 establishes the intuitively expected result that the diversity order decreases by the number $K - |\mathcal{D}|$ of users who did not decode $\mathbf{s}$ correctly. However, Lemma 3 shows that as $\gamma \to \infty$ the probability of this event behaves precisely as $\gamma^{-(K-|\mathcal{D}|)}$. These two effects annihilate each other leading to Theorem 7 that we prove next.

Proof of Theorem 7: From the theorem of total probability $P_e(\gamma) = \sum_{\mathcal{D}} P_e(\mathcal{D}) \Pr(\mathcal{D})$, and thus

$$\eta = -\lim_{\gamma\to\infty} \frac{\log\left[\sum_{\mathcal{D}} P_e(\mathcal{D}) \Pr(\mathcal{D})\right]}{\log(\gamma)} = -\min_{\mathcal{D}} \left\{ \lim_{\gamma\to\infty} \frac{\log\left[P_e(\mathcal{D}) \Pr(\mathcal{D})\right]}{\log(\gamma)} \right\}, \tag{4.19}$$

where the second equality is a manifestation of the fact that the slowest term dominates the convergence ratio. This can be further separated as

$$\eta = \min_{\mathcal{D}} \left\{ - \lim_{\gamma\to\infty} \frac{\log\left[P_e(\mathcal{D})\right]}{\log(\gamma)} - \lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathcal{D})\right]}{\log(\gamma)} \right\} = \min_{\mathcal{D}} \left\{ \eta(\mathcal{D}) - \lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathcal{D})\right]}{\log(\gamma)} \right\}. \tag{4.20}$$

The first limit in (4.20) is given by Lemma 2 and the second by Lemma 3, based on which we obtain

$$\eta = \min_{\mathcal{D}} \left\{ \max[0; \beta - (K - |\mathcal{D}|)] + (K - |\mathcal{D}|) \right\} = \beta \tag{4.21}$$

after substituting (4.17) and (4.18) into (4.20). □

The value of Theorem 7 is twofold. On the one hand, it establishes that diversity results for MISO channels carry over to judiciously designed MSC protocols. In particular, two immediate implications of Theorem 7 are stated in the following corollaries.

Corollary 3 *Diversity order of the repetition coding based MSC protocol in* [C1] *is* $\eta(\mathbf{I}, \psi_R(\cdot), \mathbf{I}) = 2$.

Proof: For repetition coding, the equivalent single-user protocol described by (4.7)-(4.15) can be readily shown to correspond to an uncoded 2×1 MISO channel which is known to provide second-order diversity. □

Corollary 4 *For the MSC protocol based on distributed ECC in* [C2] *with minimum distance* $d_{\min}$ *and code rate* R_c*, there exist matrices* $\mathbf{\Pi}_1(\mathbf{P})$ *and* $\mathbf{\Pi}_2(\mathbf{P})$ *so that*

$$\eta[\mathbf{\Pi}_2(\mathbf{P}), \psi_P(\cdot), \mathbf{\Pi}_1(\mathbf{P})] = \min(d_{\min}, \lfloor 1 + K(1 - R_c) \rfloor). \tag{4.22}$$

Proof: The result in (4.22) holds true for a single-antenna time-selective block fading channel; see e.g., [69, Ch.14]. It is thus true for MSC with distributed ECC because of Theorem 7. □

On the other hand, Theorem 7 establishes that designing good encoders $\psi(\cdot)$ is equivalent to designing diversity-enabling codes for co-located multi-antenna transmitters with the advantage that the latter is a well-understood problem. An interesting observation in this

regard is that since there are K uncorrelated Rayleigh channels in (4.15), the potential diversity order is K. MSC protocols based on distributed repetition coding or distributed ECC fail to enable this diversity order. The solution in co-located multi-antenna systems is to use complex field coding (CFC) [125], which motivates exploring CFC possibilities in a distributed setup.

4.3.2 Distributed Complex Field Coding

While in principle any matrix $\mathbf{\Phi}$ with complex entries could be used, the diversity order enabled by the DCFC protocol in [C3] depends critically on the choice of $\mathbf{\Phi}$. To make this point clear, we start by specifying the permutation matrices $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$ as KN-dimensional periodic interleavers. Letting $\mathbf{e}_i$ denote the i^{th} element of the canonical basis of $\mathbb{R}^{KN}$, we select

$$\begin{aligned}\mathbf{\Pi}_1 = \mathbf{\Pi}_{KN} &:= [\mathbf{e}_1, \mathbf{e}_N, \ldots, \mathbf{e}_{(K-1)N+1}, \mathbf{e}_2, \mathbf{e}_{N+1}, \ldots, \mathbf{e}_{KN}],\\ \mathbf{\Pi}_2 = \mathbf{\Pi}_{NK} &:= [\mathbf{e}_1, \mathbf{e}_K, \ldots, \mathbf{e}_{(N-1)K+1}, \mathbf{e}_2, \mathbf{e}_{K+1}, \ldots, \mathbf{e}_{KN}].\end{aligned} \tag{4.23}$$

The period of $\mathbf{\Pi}_1 = \mathbf{\Pi}_{KN}$ is K and consequently it changes the ordering of $\mathbf{s}$ so that in $\mathbf{r} = \mathbf{\Pi}_1\mathbf{s}$, same symbol indices across users appear consecutively in $\mathbf{r} = [s_{11}, s_{21} \ldots, s_{K1}, s_{12}, \ldots, s_{KN}]^T$. Likewise, the period of $\mathbf{\Pi}_2 = \mathbf{\Pi}_{NK}$ is N, so that $\mathbf{\Pi}_2 = \mathbf{\Pi}_1^T = \mathbf{\Pi}_1^{-1}$.

The permutation matrix $\mathbf{\Pi}_1$ models the interleaver Π_1 shown in Fig. 4.7. Each terminal in $\mathcal{D}$ correctly decodes the N symbols of all users including its own symbols that are arranged in the K vectors $\mathbf{s}_k := [s_{1k}, \ldots, s_{Nk}]^T$, $k \in [1, K]$. These KN symbols are fed to the interleaver Π_1 which outputs the N vectors $\mathbf{r}_n := [s_{n1}, \ldots, s_{nK}]^T$, $n \in [1, N]$, that contain the n^{th} symbols of all K terminals. In matrix-vector form, this relation can be written as

$$[\mathbf{r}_1^T, \ldots, \mathbf{r}_N^T]^T := \mathbf{r} = \mathbf{\Pi}_1\mathbf{s} := \mathbf{\Pi}_1[\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T. \tag{4.24}$$

We consider the DCFC based MSC protocol in which each of these $\mathbf{r}_n$ blocks is CFC-encoded independently yielding the vectors $\mathbf{u}_n := \mathbf{\Theta}\mathbf{r}_n$, corresponding to the selection $\mathbf{\Phi} := \text{diag}(\mathbf{\Theta}, \ldots, \mathbf{\Theta})$ in [C3]. These N vectors $\mathbf{u}_n := [v_{n1}, \ldots, v_{nK}]^T$, $n \in [1, N]$, are

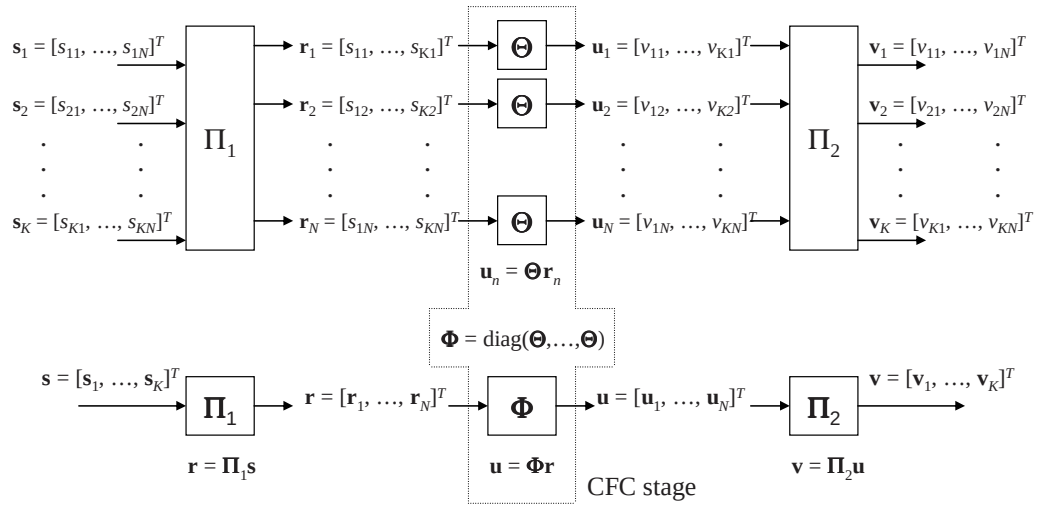

Figure 4.7: Block diagram of DCFC per cooperating user.

then fed to the interleaver Π_2 whose output consists of K vectors $\mathbf{v}_k := [v_{1k}, \ldots, v_{Nk}]^T$, $k \in [1, K]$, each containing the k^{th} element of all vectors $\mathbf{u}_n$, $n \in [1, N]$. Again, this can be written using matrix-vector notation as

$$[\mathbf{v}_1^T, \ldots, \mathbf{v}_K^T]^T := \mathbf{v} = \mathbf{\Pi}_2 \mathbf{u} := \mathbf{\Pi}_2 [\mathbf{u}_1^T, \ldots, \mathbf{u}_N^T]^T. \tag{4.25}$$

User U_k transmits the vector $\mathbf{v}_k$ leading to the phase-2 input/output relationship (4.12). We then de-interleave the received blocks at the AP to obtain [c.f. (4.12)]

$$\mathbf{\Pi}_1 \mathbf{y}_0^{(2)} = A(\mathbf{\Pi}_1 \tilde{\mathbf{D}} \mathbf{\Pi}_2)\, \mathbf{\Phi} \mathbf{r} + \mathbf{\Pi}_1 \mathbf{n}_0^{(2)}. \tag{4.26}$$

Interestingly, since $\mathbf{\Pi}_2 = \mathbf{\Pi}_1^{-1}$ we have $\mathbf{\Pi}_1 \tilde{\mathbf{D}} \mathbf{\Pi}_2 = \mathrm{diag}(h_1^{(2)}, \ldots, h_K^{(2)}, \ldots, h_1^{(2)}, \ldots, h_K^{(2)})$. Thus, upon defining $\mathbf{y}_{0n}^{(2)} := [y_{n1}^{(2)}, \ldots, y_{nK}^{(2)}]^T$, $\mathbf{n}_{0n}^{(2)} := [n_{n1}^{(2)}, \ldots, n_{nK}^{(2)}]^T$ and $\mathbf{D}_{0n}^{(2)} := \mathrm{diag}(h_1^{(2)}, \ldots, h_K^{(2)})$; and recalling that $\mathbf{\Phi}$ is block diagonal, we can write

$$\mathbf{y}_{0n}^{(2)} = A \mathbf{D}_{0n}^{(2)} \mathbf{u}_n + \mathbf{n}_{0n}^{(2)} = A \mathbf{D}_{0n}^{(2)} \mathbf{\Theta} \mathbf{r}_n + \mathbf{n}_{0n}^{(2)}, \qquad n \in [1, N]; \tag{4.27}$$

which amounts to separating (4.26) in N decoupled equations, each involving the $K \times 1$ vectors $\mathbf{r}_n$, $\mathbf{y}_{0n}^{(2)}$, and $\mathbf{n}_{0n}^{(2)}$ instead of the $KN \times 1$ vectors $\mathbf{r}$, $\mathbf{y}_0^{(2)}$, and $\mathbf{n}_0^{(2)}$.

If we finally combine $\mathbf{y}_{0n}^{(2)}$ in (4.27) with its counterpart $\mathbf{y}_{0n}^{(1)} := [y_{n1}^{(1)}, \ldots, y_{nK}^{(1)}]^T$ from (4.7) corresponding to the channel matrix $\mathbf{D}_{0n}^{(1)} := \mathrm{diag}(h_1 \ldots\ h_K)$, the ML decoder for a DCFC

based MSC protocol is

$$\hat{\mathbf{r}}_n = \arg \min_{\mathbf{r}_n \in \mathcal{S}^K} \left\| \begin{bmatrix} \mathbf{y}_{0n}^{(1)} \\ \mathbf{y}_{0n}^{(2)} \end{bmatrix}_{2K\times 1} - A \begin{bmatrix} \mathbf{D}_{0n}^{(1)} \\ \mathbf{D}_{0n}^{(2)}\boldsymbol{\Theta} \end{bmatrix}_{2K\times K} \mathbf{r}_n \right\|. \tag{4.28}$$

It is worth stressing that DCFC decoding in (4.28) operates on blocks of K symbols. (Near)-ML decoders, such as the sphere decoder [37, 68, 130], can be used to obtain $\hat{\mathbf{s}}_n$ from (4.28) with polynomial average complexity (cubic for moderate size K and SNR [37]). Certainly, if only quadratic complexity can be afforded, zero forcing, minimum mean-squared error or decision-feedback equalizer options are available but they cannot guarantee to achieve the maximum possible diversity order.

The advantage of the formulation in (4.27) is that the CFC encoder $\boldsymbol{\Theta}$ operates on $K \times 1$ symbol blocks which reduces complexity considerably relative to a KN-symbol CFC encoder. Also, basic CFC results derived for co-located multi-antenna systems [125] can be directly applied to the distributed MSC setup. In particular, it is useful to recall the notion of maximum distance separable (MDS) matrices.

Definition 3 *A matrix $\boldsymbol{\Theta}$ is called MDS with respect to the constellation $\mathcal{S}$ if and only if for any two different symbols $\mathbf{r}_1 \neq \mathbf{r}_2 \in \mathcal{S}$, all the coordinates of $\boldsymbol{\Theta}\mathbf{r}_1$ and $\boldsymbol{\Theta}\mathbf{r}_2$ are different i.e., $[\boldsymbol{\Theta}\mathbf{r}_1]_i \neq [\boldsymbol{\Theta}\mathbf{r}_2]_i$, $\forall i$.*

The MDS property leads to the following corollary of Theorem 7.

Corollary 5 *If $\boldsymbol{\Theta}$ is MDS with respect to $\mathcal{S}$, the DCFC based MSC protocol in* [C3] *with $\boldsymbol{\Phi} = \text{diag}(\boldsymbol{\Theta}, \ldots \boldsymbol{\Theta})$ and $\boldsymbol{\Pi}_1$, $\boldsymbol{\Pi}_2$ given by (4.23) enables diversity equal to the number of users; i.e.,*

$$\eta[\boldsymbol{\Pi}_2, \boldsymbol{\Phi}(\cdot), \boldsymbol{\Pi}_1] = K. \tag{4.29}$$

Proof: Because of Theorem 7 it suffices to show that $\beta[\boldsymbol{\Pi}_2, \boldsymbol{\Phi}(\cdot), \boldsymbol{\Pi}_1] = K$. Let $d_H(\mathbf{u}_{n1}, \mathbf{u}_{n2})$ be the Hamming distance between codewords $\mathbf{u}_{n1}$ and $\mathbf{u}_{n2}$. When $\mathbf{D}_{0n}^{(2)}$ is Rayleigh distributed, the diversity order is $\beta[\boldsymbol{\Pi}_2, \boldsymbol{\Phi}(\cdot), \boldsymbol{\Pi}_1] = \min_{\mathbf{u}_{n1}, \mathbf{u}_{n2}} d_H(\mathbf{u}_{n1}, \mathbf{u}_{n2})$ [60, Sec. 2.1.2]. The MDS property guarantees that $\min_{\mathbf{u}_{n1}, \mathbf{u}_{n2}}[d_H(\mathbf{u}_{n1}, \mathbf{u}_{n2})] = K$ for $\mathbf{u}_n = \boldsymbol{\Theta}\mathbf{r}_n$ and consequently $\beta[\boldsymbol{\Pi}_2, \boldsymbol{\Phi}(\cdot), \boldsymbol{\Pi}_1] = K$. □

Remark 11 Relative to repetition based single-source cooperation (SSC), the MSC protocol based on distributed ECC or CFC can also enhance coding gains because relay transmissions are coded across time and space. As each source in MSC is served by multiple cooperators, for the same spectral efficiency, even ECC based MSC can achieve higher diversity gains than SSC. And since each cooperator serves multiple sources simultaneously, for the same diversity order, MSC can offer higher spectral efficiency than SSC. With regards to the cooperative multi-user protocol in [76] which relies on the presence of idle users, the unifying MSC protocol here does not require idle users. However, if K_I idle users are available and willing to cooperate, the protocol here can also take advantage of them to increase the diversity order up to $K + K_I$ per cluster.

4.4 Multi-cluster operation

Let us now return to the multi-cluster setting with $L > 1$ non-overlapping clusters communicating with the AP. Cluster separation can be accomplished with any multiple access (MA) scheme which in principle should not affect the properties of the DCFC based MSC protocol. However, it will turn out that MA with CFC can affect spectral efficiency of the MSC protocol.

Recall that $\mathbf{s}_{lk} := [s_{lk1}, \ldots, s_{lkN}]^T$ denotes the data packet of U_{lk} and $\mathbf{s}_l := [\mathbf{s}_{l1}^T, \ldots, \mathbf{s}_{lK}^T]^T$ the l^{th} cluster's aggregate block[1]. To separate clusters at the AP we rely on CDMA with spreading code signatures $\{c_l(t)\}_{l=1}^L$. To this end, the waveform transmitted by U_{lk} is

$$x_{lk}^{(1)}(t) = A \sum_{n=1}^{N} s_{lkn} c_l[t - (N(k-1)+n)T_s], \qquad t \in [0, kNT_s). \tag{4.30}$$

The correlation between $c_l(t)$ and $c_m(t)$ will be denoted as $\rho_{ml} := \int_0^{T_s} c_l(t)c_m(t)dt$ and arranged in the symmetric $L \times L$ matrix $\mathbf{R}$ with entries $[\mathbf{R}]_{ml} := \rho_{ml}$.

User U_{lk} receives the superposition of what the remaining $LK-1$ users transmit, namely

$$y_{lk}^{(1)}(t) = \sum_{m=1}^{L} \sum_{j=1}^{K} h_{lk,mj}\, x_{mj}^{(1)}(t) + n_{lk}^{(1)}(t). \tag{4.31}$$

[1] Although generalizations are immediate, to avoid further complication of the already heavy notation we will assume that the clusters have equal number of active users; i.e., $K_l = K$, $\forall l \in [1, L]$.

For notational simplicity, let us consider the waveform $y_{00}^{(1)}(t) = y^{(1)}(t)$ received by the AP. Notice that depending on the correlation matrix $\mathbf{R}$, optimal reception may require *joint* detection of the $L \times 1$ vector $\bar{\mathbf{s}}_{kn} := [s_{1kn}, \ldots, s_{Lkn}]^T$ containing the n^{th} symbol of the k^{th} user across the L clusters. The joint detector will rely on the $L \times 1$ decision vector $\mathbf{y}_{kn}^{(1)} := [y_{1kn}^{(1)}, \ldots, y_{Lkn}^{(1)}]^T$ whose components are given by

$$y_{lkn}^{(1)} := \int_{kn(T_s-1)}^{knT_s} y_{lk}^{(1)}(t) c_l(t - knT_s) dt = A \sum_{m=1}^{L} \rho_{ml} h_{mk} s_{mkn} + n_{lkn}^{(1)}. \tag{4.32}$$

Upon defining $\mathbf{D}_k^{(1)} := \text{diag}(h_{k1}, \ldots, h_{kL})$, we can rewrite (4.32) in matrix-vector form as

$$\mathbf{y}_{kn}^{(1)} = A\mathbf{R}\mathbf{D}_k^{(1)} \bar{\mathbf{s}}_{kn} + \mathbf{n}_{kn}^{(1)}, \qquad k \in [1, K],\ n \in [1, N]. \tag{4.33}$$

The complexity of ML detection required to recover $\bar{\mathbf{s}}_{kn}$ from $\mathbf{y}_{kn}^{(1)}$ depends on the dimensionality L. While (4.33) models reception at the AP, a similar relationship characterizes reception in every cooperating user allowing U_{lk} to construct the estimate $\hat{\mathbf{s}}_{lk}$ of its cluster's aggregate block $\mathbf{s}_l$. Similar to Section 4.3, we define $\mathcal{D}_l := \{U_{lk} \mid \hat{\mathbf{s}}_{lk} = \mathbf{s}_l\} \subseteq \mathcal{U}_l$; and let $h_{lk}^{(2)} = h_{lk}$ if $U_{lk} \in \mathcal{D}_l$, and $h_{lk}^{(2)} = 0$ else. As before, users $U_{lk} \in \mathcal{D}_l$ participate in phase-2.

Each cluster in phase-2 operates separately, repeating the steps described in Section 4.3.2 to construct $\mathbf{v}_l := [v_{l11}, \ldots, v_{l1N}, v_{l21}, \ldots, v_{lKN}]^T = \mathbf{\Pi}_2 \mathbf{\Phi} \mathbf{\Pi}_1 \mathbf{s}_l$ with $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ as in (4.23) and $\mathbf{\Phi} = \text{diag}(\mathbf{\Theta}, \ldots, \mathbf{\Theta}) \in \mathbb{R}^{NK \times NK}$. Each user $U_{lk} \in \mathcal{D}_l$ then transmits the sub-block $\mathbf{v}_{lk} := [v_{lk1}, \ldots, v_{lkN}]^T$. Except for notation, these steps are identical to those in Section 4.3.2. The difference is in the received signal which now comprises the superposition of waveforms transmitted from users in all L clusters

$$y_0^{(2)}(t) = A \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{n=1}^{N} v_{lkn} c_l[t - (N(k-1) + n)T_s] + n_0^{(2)}(t). \tag{4.34}$$

As in (4.32), we let $y_{lkn}^{(2)} := \int_{kn(T_s-1)}^{knT_s} y_0^{(2)}(t) c_l(t - knT_s) dt$ so that upon defining $\mathbf{y}_{kn}^{(2)} := [y_{1kn}^{(2)}, \ldots, y_{Lkn}^{(2)}]^T$, $\bar{\mathbf{v}}_{kn} := [v_{1kn}, \ldots, v_{Lkn}]^T$ and $\mathbf{D}_{0k}^{(2)} = \text{diag}(h_{1k}^{(2)}, \ldots, h_{Lk}^{(2)})$, we can write $\mathbf{y}_{kn}^{(2)} = A\mathbf{R}\mathbf{D}_{0k}^{(2)} \bar{\mathbf{v}}_{kn} + \mathbf{n}_{kn}^{(2)}$, the counterpart of (4.33) for phase-2. For future use, it is convenient to define $\mathbf{y}_{0n}^{(2)} := [{\mathbf{y}^{(2)}}_{1n}^T, \ldots, {\mathbf{y}^{(2)}}_{Kn}^T]^T$, $\bar{\mathbf{v}}_n := [\bar{\mathbf{v}}_{1n}^T, \ldots, \bar{\mathbf{v}}_{Kn}^T]^T$, $\bar{\mathbf{R}} := \text{diag}(\mathbf{R}, \ldots, \mathbf{R})$ and $\tilde{\mathbf{D}} := \text{diag}(\mathbf{D}_{01}^{(2)}, \ldots, \mathbf{D}_{0K}^{(2)})$ and write

$$\mathbf{y}_{0n}^{(2)} = A\bar{\mathbf{R}}\tilde{\mathbf{D}}\bar{\mathbf{v}}_n + \mathbf{n}_{0n}^{(2)}. \tag{4.35}$$

On the other hand, let $\mathbf{u}_{ln} := [u_{l1n}, \ldots, u_{lKn}]^T$, $\mathbf{u}_n := [\mathbf{u}_{1n}^T, \ldots, \mathbf{u}_{Ln}^T]^T$ and $\mathbf{\Pi}_{LK} := [\mathbf{e}_1, \mathbf{e}_L, \ldots, \mathbf{e}_{(L-1)K+1}, \mathbf{e}_2, \mathbf{e}_{K+1}, \ldots, \mathbf{e}_{KL}]$ be a KL-dimensional periodic interleaver with period L. According to these definitions, we have $\bar{\mathbf{v}}_n = \mathbf{\Pi}_{LK}\mathbf{u}_n$. Also, note that since $\mathbf{u}_{ln} = \mathbf{\Theta}\mathbf{r}_{ln}$, for $\mathbf{r}_n := [\mathbf{r}_{1n}^T, \ldots, \mathbf{r}_{Ln}^T]^T$ and $\bar{\mathbf{\Phi}} = \mathrm{diag}(\mathbf{\Theta}, \ldots, \mathbf{\Theta}) \in \mathbb{R}^{LK \times LK}$, we have that

$$\mathbf{y}_{0n}^{(2)} = A\bar{\mathbf{R}}\tilde{\mathbf{D}}\mathbf{\Pi}_{LK}\bar{\mathbf{\Phi}}\mathbf{r}_n + \mathbf{n}_{0n}^{(2)}. \tag{4.36}$$

Concatenating (4.33) and (4.36) we obtain the ML decoder for $\mathbf{r}_n$ as

$$\hat{\mathbf{r}}_n = \arg \min_{\mathbf{r}_n \in \mathcal{S}^{LK}} \left\| \begin{bmatrix} \mathbf{y}_{0n}^{(1)} \\ \mathbf{y}_{0n}^{(2)} \end{bmatrix}_{2LK \times 1} - A \begin{bmatrix} \bar{\mathbf{R}}\mathbf{D}_0^{(1)} \\ \bar{\mathbf{R}}\tilde{\mathbf{D}}\mathbf{\Pi}_{LK}\bar{\mathbf{\Phi}} \end{bmatrix}_{2LK \times LK} \mathbf{r}_n \right\|. \tag{4.37}$$

Dimensionality of the multi-cluster ML decoder (4.37) is KL that has to be compared with K, the corresponding dimensionality of the ML decoder in (4.28) for the single-cluster case.

Even though the input-output relationships (4.33) and (4.36) as well as (4.7) and (4.12) model different systems they exhibit similar forms. An important consequence of this observation is that Corollary 3 establishing the diversity order of a single-cluster DCFC based MSC protocol can be readily generalized.

Corollary 6 *If $\mathbf{\Theta}$ is MDS with respect to $\mathcal{S}$, the multi-cluster DCFC based MSC protocol with the ML decoder in (4.37) achieves diversity equal to the number of users in each cluster; i.e.,*

$$\eta[\mathbf{\Pi}_2, \bar{\mathbf{\Phi}}, \mathbf{\Pi}_1, \mathbf{R}] = K. \tag{4.38}$$

Proof: For a Rayleigh channel, the coefficients h_{lk} are complex Gaussian and consequently the channel $\mathbf{D}_{0eq}^{(1)} := \bar{\mathbf{R}}\mathbf{D}_0^{(1)}$ is also Rayleigh. As in Corollary 5, notice from (4.35) that the diversity for Rayleigh distributed channel $\mathbf{D}_{0eq}^{(2)} := \bar{\mathbf{R}}\tilde{\mathbf{D}}$ is $\beta[\mathbf{\Pi}_2, \bar{\mathbf{\Phi}}, \mathbf{\Pi}_1, \mathbf{R}] = \min_{\bar{\mathbf{v}}_{n1}, \bar{\mathbf{v}}_{n2}} d_H(\bar{\mathbf{v}}_{n1}, \bar{\mathbf{v}}_{n2})$, [60, Sec. 2.1.2]. Since $\mathbf{\Pi}_{LK}$ is a permutation matrix, it follows that $d_H(\bar{\mathbf{v}}_{n1}, \bar{\mathbf{v}}_{n2}) = d_H(\bar{\mathbf{\Phi}}\mathbf{r}_{n1}, \bar{\mathbf{\Phi}}\mathbf{r}_{n2})$. The minimum d_H is achieved when all but one cluster transmit the same symbol block. Supposing without loss of generality that the first cluster transmits the distinct symbol block we have that $\mathbf{r}_{n1} := [\mathbf{r}_{1n1}^T, \mathbf{r}_{2n}^T \ldots, \mathbf{r}_{Ln}^T]^T$ and $\mathbf{r}_{n2} := [\mathbf{r}_{1n2}^T, \mathbf{r}_{2n}^T \ldots, \mathbf{r}_{Ln}^T]^T$. The MDS property guarantees that for these $\mathbf{r}_{n1}, \mathbf{r}_{n2}$ selected,

the minimum Hamming distance is $d_H(\bar{\mathbf{\Phi}}\mathbf{r}_{n1}, \bar{\mathbf{\Phi}}\mathbf{r}_{n2}) = K$. Invoking now Theorem 7, we deduce that $\eta[\mathbf{\Pi}_2, \bar{\mathbf{\Phi}}, \mathbf{\Pi}_1, \mathbf{R}] = \beta[\mathbf{\Pi}_2, \bar{\mathbf{\Phi}}, \mathbf{\Pi}_1, \mathbf{R}] = d_H(\bar{\mathbf{\Phi}}\mathbf{r}_{n1}, \bar{\mathbf{\Phi}}\mathbf{r}_{n2}) = K$. □

As expected, Corollary 4 proves that the diversity enabled by DCFC remains invariant regardless of the structure of the correlation matrix $\mathbf{R}$.

Remark 12 For clarity, we have considered MA in fixed cooperative networks. However, the same scheme and results are also valid for ad-hoc networks. In this case, different sources per cluster cooperate while communicating with (possibly) different destinations. Different clusters operate without coordination.

4.4.1 Effect of under-spreading in spectral efficiency

Consider a set of orthonormal functions $\mathcal{N} := \{\nu_s(t)\}_{s=1}^{S}$ with $\int_0^{T_s} \nu_{s_1}(t)\nu_{s_2}(t)dt = \delta(s_1 - s_2)$, where $\delta(\cdot)$ denotes Kronecker's delta. It is customary to write the signature waveforms in (4.30) as the linear combination

$$c_l(t) = \sum_{s=1}^{S} c_{ls}\nu_s(t), \qquad t \in [0, T_s),\ l \in [1, L], \tag{4.39}$$

where the vector $\mathbf{c}_l := [c_{l1}, \ldots, c_{lS}]^T$ is the spreading code specific to the cluster $\mathcal{U}_l$. Arranging the codes in a matrix $\mathbf{C} := [\mathbf{c}_1, \ldots, \mathbf{c}_L]$ we can write the correlation matrix as $\mathbf{R} = \mathbf{C}^{\mathcal{H}}\mathbf{C}$. Changing the set $\mathcal{N}$ we can model different CDMA systems; if the functions are delayed versions of each other $\nu_s(t) = p[St - (s-1)T_s)]$, then (4.39) amounts to symbol-periodic direct sequence (DS)-CDMA; if they are different subcarriers, $\nu_s(t) = \exp[j2\pi(s-1)t/T_s]p(t)$, then (4.39) models multi-carrier (MC)-CDMA.

Transmission of $\nu_s(t)$ requires S times more bandwidth than transmission of the pulses $p(t)$ in (4.6). Consequently, the spectral efficiency of multi-cluster DCFC is

$$\xi = L/(2S). \tag{4.40}$$

An important choice in the selection of $\mathbf{C}$ is whether the spreading gain S constrains *a fortiori* the number of codes L or not. This calls for distinguishing between *under-spread* and *over-spread* MA:

Definition 4 *In an over-spread MA system, the number of codes L and the spreading gain S are constrained by $S \geq L$. We say that an MA system is under-spread if L and S can be selected independently.*

Over-spread orthonormal MA: In this case, $\mathbf{C}$ is formed by orthonormal vectors, e.g., Walsh-Hadamard sequences, so that $\mathbf{R} := \mathbf{C}^{\mathcal{H}}\mathbf{C} = \mathbf{I}_L$. But the latter requires $L \leq S$ because a set of orthonormal vectors in $\mathbb{R}^S$ cannot contain more than S elements. Thus, orthonormal MA is over-spread in the sense of Definition 4.

Under-spread MA: Symbol-periodic non-orthogonal signatures, including those in MC-CDMA and DS-CDMA with Gold or Kasami sequences [24], implement under-spread MA since L can be much larger than S. Long pseudo-noise (PN) sequences also give rise to under-spread MA with approximately uncorrelated signatures. Since the latter can be theoretically infinite, L and S are decoupled and long code DS- or MC-CDMA is also under-spread in the sense of Definition 4.

Since in orthonormal MA, clusters do not interfere with each other, the multi-cluster model in (4.37) can be reduced to a set of single-cluster models (4.28). Thus, the ML decoding space dimension is reduced from KL to K. PN sequences, on the other hand, have found widespread use due to their robustness to propagation delays and relaxed synchronization requirements. The decision as to whether to use under- or over-spread MA may also depend on other factors as well.

MSC protocols with under-spread versus over-spread MA are fundamentally different in terms of bandwidth efficiency. In over-spread MA the spectral efficiency of MSC protocols is hard limited by $\xi_{\text{MSC}} \leq 1/2$ [c.f. (4.40) and Definition 4] and cooperation comes at the price of reducing the spectral efficiency $\xi_{\text{NC}} = 1$ of the corresponding non-cooperative system. In e.g., orthonormal MA, this is because $L \leq S$ clusters can have orthogonal signatures. In under-spread MA, bandwidth efficiency and cooperative diversity are not necessarily traded off since L and S are decoupled. Indeed, we can obtain $\xi_{\text{MSC}} = \xi_{NC}$ by reducing the spreading gain by half, i.e., $S_{\text{MSC}} = S_{NC}/2$ while maintaining the same number of clusters L [c.f. (4.40)]. Note that even if we reduce the spreading gain by half, after completing both

Table 4.1: Comparison of different protocols

	Performance metrics	Repetition coding	Distributed ECC	DCFC	Non-cooperative
over-spread MA	diversity (η)	2	$\min(d_{\min}, \lfloor 1 + K(1 - R_c) \rfloor)$	K	1
	spectral eff. (ξ)	1/2	1/2	1/2	1
	complexity	1	KN	K	1
under-spread MA	diversity (η)	2	$\min(d_{\min}, \lfloor 1 + K(1 - R_c) \rfloor)$	K	1
	spectral eff. (ξ)	1	1	1	1
	complexity	L	LKN	LK	1

MSC phases each information symbol has been transmitted twice and the effective coding gain is still the same as in non-cooperative MA. Nonetheless, a consequence of Corollary 6 is that the diversity gain is $\eta[\mathbf{\Pi}_2, \bar{\mathbf{\Phi}}, \mathbf{\Pi}_1, \mathbf{R}] = K$, regardless of the correlation structure $\mathbf{R}$. Thus, under-spread MA with DCFC *achieves full diversity* without sacrificing spectral efficiency.

4.5 Comparing MSC with non-cooperative protocols

So far, we have considered three different MSC protocols, namely distributed repetition coding, distributed ECC and DCFC defined in [C1], [C2] and [C3], respectively. We also distinguished between under- and over-spreading for cluster separation as per Definition 4, for a total of six different alternatives. These alternatives differ in their diversity η [c.f. (4.1)], spectral efficiency ξ [c.f. (4.40)] and decoding complexity as we summarize in Table 4.1.

Repetition coding can afford the lowest decoding complexity, but also enables the smallest diversity order. Moreover, the diversity order it enables is independent of the number of users in the cluster. The diversity order can be increased with either distributed ECC or DCFC at the expense of increasing decoding complexity. It is known that $\eta \approx 4$ brings the wireless channel within a 10% of an AWGN channel's error performance [121], meaning that $K \approx 4$ captures enough of the diversity advantage. Thus, a slight complexity increase brings in a substantial error performance gain. This is particularly true for DCFC that achieves full; i.e., $\eta = K$, diversity. For distributed ECC a larger cluster is possibly needed.

Under- and over-spreading are fundamentally different in terms of bandwidth efficiency

ξ. In the rows corresponding to over-spread MA, ξ is reduced from 1 to 1/2 for any of the MSC protocols. The value of ξ in the corresponding columns of Table 4.1 for under-spread MA is not affected when we move from non-cooperative MA to MSC. The value $\xi = 1$ is an arbitrary selection and should be interpreted as an option to allow for a fair comparison between over-spread non-cooperative MA and under-spread cooperative MA. Interestingly, the use of under-spread MA with DCFC achieves full diversity K while avoiding the bandwidth penalty usually associated with cooperative protocols as we can see by comparing the second with the fifth row of Table 4.1.

All in all, in a complexity-limited system repetition coding offers the best MSC protocol, while in a bandwidth-limited setup DCFC-based MSC with under-spreading for cluster separation should be preferred. In intermediate cases, DCFC-based MSC with (over-spread) orthonormal cluster separation achieves full diversity with reasonable spectral-efficiency ($\xi = 1/2$) and a modest increase in complexity.

4.6 Simulations

In this section, we present simulated examples to corroborate our analytical claims. Each user transmits blocks with $N = 50$ symbols per TDMA slot. Except for one example, we assume error-free channels between users. We choose the CFC encoder $\mathbf{\Theta}$ to be the unitary Vandermonde matrix in [125]

$$\mathbf{\Theta} = \frac{1}{\sqrt{K}}\mathbf{F}_K^H \text{diag}(1, \alpha, \ldots, \alpha^{K-1}), \tag{4.41}$$

where $\mathbf{F}_K$ is the $K \times K$ fast Fourier transform (FFT) matrix with (i,j)th entry $[\mathbf{F}_K]_{ij} := e^{-j2\pi(i-1)(j-1)/K}$; and $\alpha := e^{j\pi/(2K)}$ if K is power of 2, $\alpha := e^{j\pi/9}$ if $K = 3$, and $\alpha := e^{j\pi/25}$ if $K = 5$.

DCFC based MSC with orthonormal MA.

Consider first the DCFC based MSC protocol with orthonormal CDMA signatures used for cluster separation. According to Table 4.1, the ML decoder operates on blocks of length K, the spectral efficiency is $\xi = 1/2$ and the diversity order is $\eta = K$. To benchmark performance consider error-free links between users, in which case $\mathcal{D} \equiv \mathcal{U}$. Fig. 4.8 demonstrates

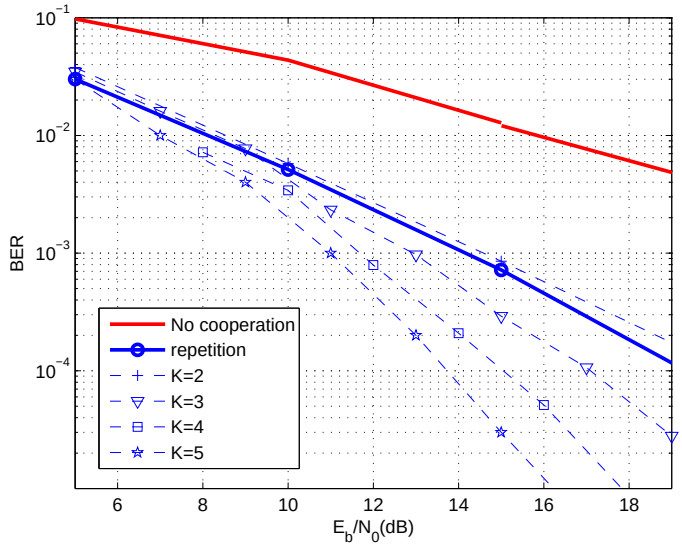

Figure 4.8: BER of orthonormal DCFC-based MSC with variable number of users, and error-free user-to-user links.

how the bit error rate (BER) varies with K for a DCFC based MSC protocol. We verify that the diversity order is, indeed, equal to the number of users K. For reference, we also depict the BER of a non-cooperative system and repetition based MSC [96]. For $K = 2$ repetition based MSC outperforms DCFC based MSC by a small margin. This is because in this case both protocols have the same diversity gain but the coding gain of DFCF is smaller. The advantage of DCFC is apparent for larger cooperating clusters. Setting e.g., $K = 5$, we can see that with a minimal investment in decoding complexity, DCFC based MSC returns a $4 - 5$ dB gain with respect to repetition based MSC due to the increase in diversity from $\eta = 2$ to $\eta = K = 5$.

Even though the simulated curves in Fig. 4.8 are for error-free user-to-user links, the same results are obtained when we account for the effect of decoding errors in these links as verified by Fig. 4.9 for $K = 3$. We consider different values of the relative SNR $\Delta := \gamma_{k,j}/\gamma_k = \mathrm{E}[h_{k,j}^2]/\mathrm{E}[h_k^2]$, where we recall $\gamma_{k,j}$ is the average SNR of the $U_k \rightarrow U_j$ link and γ_k is the average SNR in the $U_k \rightarrow$AP link. Regardless of Δ, the diversity order is

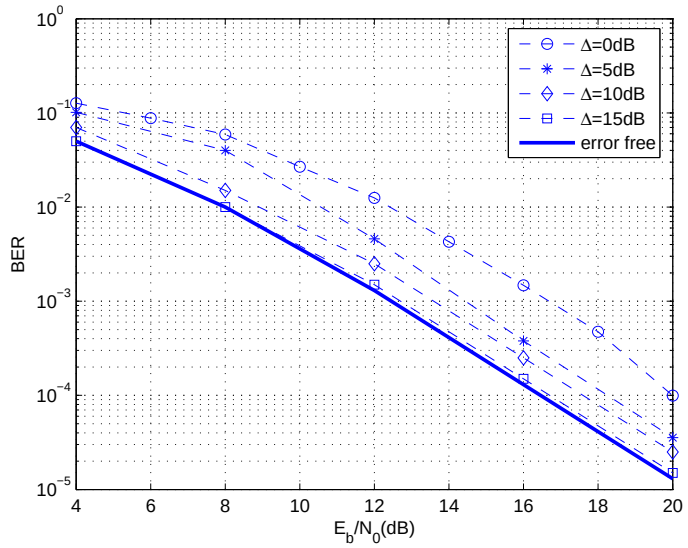

Figure 4.9: BER of orthonormal DCFC-based MSC with variable relative SNRs in the links between user pairs.

always $\eta = K = 3$ as asserted by Theorem 7, but the coding gain changes, as predicted. Notwithstanding, the gap between error-free $U_k \to U_j$ links and $\Delta = 5$dB is approximately 2dB, and reduces to less than 1dB for $\Delta = 10$dB. Thus, in many practical settings proximity of cooperators ensures that MSC protocols work almost as well as non-cooperative single-user multi-antenna systems.

DCFC based MSC with under-spread MA.

Spectral efficiency in the simulated systems of the previous subsection is $\xi = 1/2$. This is not the case for DCFC with under-spread MA which according to Table 4.1 requires ML decoding on blocks of length KL, but attains spectral efficiency $\xi = 1$ and diversity order $\eta = K$. In Fig. 4.10 we show BER for $K = 3$, $S = 8$ and $L = 4$ – $L = 8$ with PN codes used to implement under-spread cluster separation. Verifying Corollary 6, the diversity order is $\eta = K = 3$ regardless of the number of clusters L. When $L = 8$ the spectral efficiency is $\xi = 1$ and it is pertinent to compare DCFC with a non-cooperative protocol with orthonormal MA (for which $\xi = 1$ too). The diversity enabled by DCFC

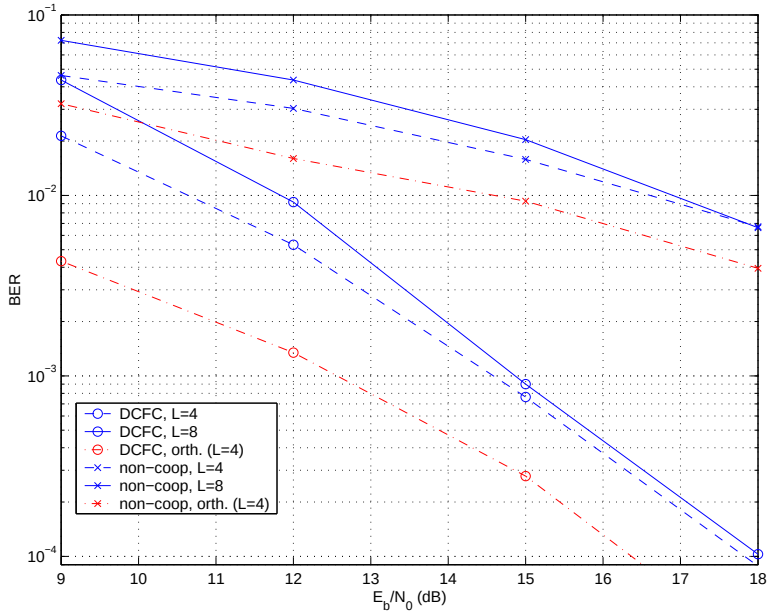

Figure 4.10: BER of under-spread DCFC-based MSC with different values of spectral efficiency.

leads to a considerable BER reduction. When $L = 4$ the spectral efficiency is $\xi = 1/2$. In this case, it is possible to use DCFC based MSC with orthonormal MA. We can see that gaining in spectral efficiency with DCFC entails a loss in coding gain of about 2dB. Interestingly, the coding gain is affected by the use of under-spread MA but the diversity order is not. Complexity allowing, DCFC based MSC with under-spread MA is the choice for bandwidth-limited scenarios, whereas if bandwidth is plenty the use of orthonormal MA should be preferred for its larger coding gain.

DCFC versus Distributed ECC.

Even though we established that the diversity order of DCFC is in general larger than the diversity of distributed (D)ECC in [C2] (see Table 4.1), the latter has in general a larger coding gain. To demonstrate these differences, we consider clusters with $K = 3$ users and compare DCFC based MSC against MSC based on distributed convolutional coding (DCC) with rate 1/2 and generator in octal form [15/7]. The free distance of this code is $d_{\min} = 5$

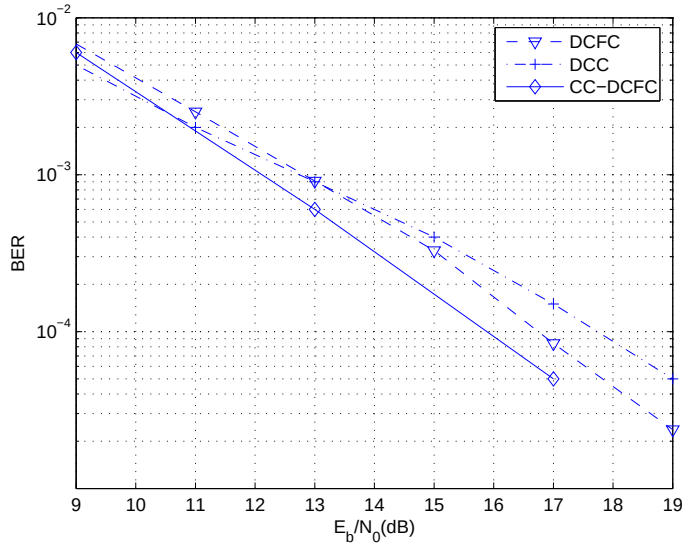

Figure 4.11: BER of orthonormal DCFC and DCC based MSC protocols.

and consequently the diversity order is $\eta_{DCC} = 2$. The interleavers Π_1, Π_2 are designed as in [117] to ensure that this diversity is actually achieved.

Fig. 4.11 illustrates that due to its larger diversity order DCFC outperforms DCC at high SNR while the opposite is true at low SNR due to the larger coding gain of DCC based MSC. We can also use CC and DCFC together to jointly exploit the coding gain of CC and the diversity order of DCFC. This is done by encoding each user's bits with a CC *before* transmitting $\mathbf{s}_k$. At the receiver, we adopt soft iterative decoding between CC and DCFC decoders along the lines of [125]. As depicted in Fig. 4.11, there is about 1dB gain achieved with a CC of memory 2. Another advantage of DCFC is that the interleavers are simple periodic multiplexers as opposed to carefully designed interleavers needed for DCC in [117].

4.7 Summary

We introduced a general multi-source cooperation (MSC) framework for multi-cluster networks that allows flexible tradeoffs between error performance, spectral efficiency and complexity. Our MSC protocols rely on code division multiple access (CDMA) for cluster

separation and time division multiple access (TDMA) for user separation per cluster. The unifying framework includes many existing protocols as particular cases and suggests the introduction of distributed complex field coding (DCFC) to enable diversity as high as the number of per-cluster users. We also demonstrated the different spectral efficiency properties of over-spread, e.g., orthonormal, and under-spread, e.g., MC-CDMA or DS-CDMA with Gold or Kasami sequences, cluster separation. Whereas in the former cooperative diversity is traded off for bandwidth in the latter there is no bandwidth penalty associated with user cooperation.

Adjusting the number of cooperating users, MSC encoder, spreading gain and/or number of clusters, our general MSC framework is flexible to tradeoff among spectral-efficiency, decoding complexity and diversity. By increasing complexity, the combination of DCFC with under-spread multiple access enables high order diversity with spectral efficiency up to that of non-cooperative systems. In cases where bandwidth is not the limiting resource, DCFC-based MSC with over-spread orthonormal cluster separation allows one to collect full diversity with reasonable spectral-efficiency and a modest increase in complexity.

Appendices

4.7.1 Proof of Lemma 2

Let $\delta(\mathbf{f};\mathbf{g})$ be the indicator function of $\mathbf{f} \neq \mathbf{g}$ taking on the values $\delta(\mathbf{f};\mathbf{g}) = 1$ if $\mathbf{f} \neq \mathbf{g}$ and $\delta(\mathbf{f};\mathbf{g}) = 0$ if $\mathbf{f} = \mathbf{g}$; and consider two distinct codewords $[\mathbf{s}^T, \mathbf{v}^T] = [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_K^T]$ and $[\tilde{\mathbf{s}}^T, \tilde{\mathbf{v}}^T] = [\tilde{\mathbf{s}}_1^T, \ldots, \tilde{\mathbf{s}}_K^T, \tilde{\mathbf{v}}_1^T, \ldots, \tilde{\mathbf{v}}_K^T]$. If $\mathcal{D} = \mathcal{U}$, then MSC is equivalent to a multi-antenna channel, in which case the diversity achieved depends on the triplet $(\mathbf{\Pi}_2, \psi(\cdot), \mathbf{\Pi}_1)$. If we define $\beta(\mathbf{s};\tilde{\mathbf{s}})$ as the pairwise error probability (PEP) diversity order we have [19]

$$\beta(\mathbf{s};\tilde{\mathbf{s}}) := -\lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathbf{s}\to\tilde{\mathbf{s}}|\mathcal{D}=\mathcal{U})\right]}{\log(\gamma)} = \sum_{k=1}^{K} \delta\left([\mathbf{s}_k^T, \mathbf{v}_k^T]; [\tilde{\mathbf{s}}_k^T, \tilde{\mathbf{v}}_k^T]\right). \tag{4.42}$$

That is, the probability that we declare $\tilde{\mathbf{s}}$ when the actual transmitted block is $\mathbf{s}$ goes to zero as $\gamma^{-\beta(\mathbf{s};\tilde{\mathbf{s}})}$, with $\beta(\mathbf{s};\tilde{\mathbf{s}})$ given by (4.42). Consequently, if $[\mathbf{s}_k^T, \mathbf{v}_k^T] = [\tilde{\mathbf{s}}_k^T, \tilde{\mathbf{v}}_k^T]$ user U_k does not contribute to the diversity order of this particular pair and if $[\mathbf{s}_k^T, \mathbf{v}_k^T] \neq [\tilde{\mathbf{s}}_k^T, \tilde{\mathbf{v}}_k^T]$ U_k contributes one unit to the PEP exponent in (4.42).

If $U_k \notin \mathcal{D}$ then $h_k^{(2)} = 0$, which is equivalent to having $\mathbf{v}_k$ and $\tilde{\mathbf{v}}_k$ punctured. Thus, each $U_k \notin \mathcal{D}$ reduces $\beta(\mathbf{s};\tilde{\mathbf{s}})$ by (at most) 1 which implies that the PEP diversity order conditioned on $\mathcal{D}$ is bounded as

$$\begin{aligned}\eta(\mathbf{s};\tilde{\mathbf{s}}|\mathcal{D}) &:= -\lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathbf{s}\to\tilde{\mathbf{s}}|\mathcal{D})\right]}{\log(\gamma)} = \sum_{k|U_k\in\mathcal{D}} \delta\left([\mathbf{s}_k^T, \mathbf{v}_k^T]; [\tilde{\mathbf{s}}_k^T, \tilde{\mathbf{v}}_k^T]\right) \\ &\geq \max[0; \beta(\mathbf{s};\tilde{\mathbf{s}}) - (K - |\mathcal{D}|)]. \end{aligned} \tag{4.43}$$

Finally, note that $\beta = \min_{\mathbf{s};\tilde{\mathbf{s}}} \beta(\mathbf{s};\tilde{\mathbf{s}})$ and likewise for $\eta(\mathcal{D})$. But if (4.42) holds for any pair of codewords it must hold for their minima and (4.17) follows. □

4.7.2 Proof of Lemma 3

Let $F(k,j) := \{\hat{\mathbf{s}}_{k,j} \neq \mathbf{s}_j\}$ denote the event that U_k fails to correctly decode U_j's message. The probability of $F(k,j)$ can be obtained by averaging over the realizations of $h_{k,j}$ to obtain

$$\Pr\{F(k,j)\} = \mathrm{E}_{h_{k,j}}[\Pr\{F(k,j)|h_{k,j}\})] = \mathrm{E}_{h_{k,j}}\left[1 - \left(1 - Q\left(\sqrt{\kappa\gamma|h_{k,j}|^2}\right)\right)^N\right], \tag{4.44}$$

where $\gamma|h_{k,j}|^2$ is the instantaneous SNR in the $U_i \to U_j$ link, N is the length of the symbol vector from one user, and κ is a constant dependent on the constellation. It is not difficult to show that as γ increases we have, see e.g., [69, Chap. 14]

$$\lim_{\gamma\to\infty} \frac{\log\left[\Pr\{F(k,j)\}\right]}{\log(\gamma)} = \lim_{\gamma\to\infty} \frac{\log\left[\mathrm{E}_{h_{k,j}}\left(NQ\left(\sqrt{\kappa\gamma|h_{k,j}|^2}\right)\right)\right]}{\log(\gamma)} = -1. \tag{4.45}$$

Since the events $F(k,j)$ are independent, the probability of a given user being part of $\mathcal{D}$ is such that

$$\Pr(U_k \notin \mathcal{D}) \leq \Pr\left(\bigcup_{j=1,j\neq k}^{K} F(k,j)\right) = \sum_{j=1,j\neq k}^{K} \Pr\left(F(k,j)\right). \tag{4.46}$$

But since (4.45) is valid for all $F(k,j)$ we have that [c.f. (4.45), (4.46)]

$$\lim_{\gamma\to\infty} \frac{\log\left[\Pr(U_k \notin \mathcal{D})\right]}{\log(\gamma)} \leq \lim_{\gamma\to\infty} \frac{\log\left[\sum_{j=1,j\neq k}^{K} \Pr\left(F(k,j)\right)\right]}{\log(\gamma)} = -1. \tag{4.47}$$

On the other hand, the probability of $\mathcal{D}$ can be bounded as

$$\Pr(\mathcal{D}) = \prod_{U_k\notin\mathcal{D}} \Pr(U_k \notin \mathcal{D}) \prod_{U_k\in\mathcal{D}} \Pr(U_k \in \mathcal{D}) \leq \prod_{U_k\notin\mathcal{D}} \Pr(U_k \notin \mathcal{D}), \tag{4.48}$$

where the inequality follows since $\Pr(U_k \in \mathcal{D}) \leq 1$. Using (4.48) we can finally write for the limit in (4.18)

$$\lim_{\gamma\to\infty} \frac{\log\left[\Pr(\mathcal{D})\right]}{\log(\gamma)} = \sum_{U_k\notin\mathcal{D}} \lim_{\gamma\to\infty} \frac{\log\left[\Pr(U_k \notin \mathcal{D})\right]}{\log(\gamma)}. \tag{4.49}$$

Since the sum in (4.49) contains $K - |\mathcal{D}|$ elements, (4.18) follows after substituting (4.47) into (4.49). □

Chapter 5

Cooperative diversity in random access networks

Instead of agreeing on a fixed channel allocation, RA networks let users transmit at random contending to reach the common AP. Letting users transmit packets independently with probability p implies that successful packet delivery depends not only on the physical channel but on how many other users decided to transmit, leading to a packet delivery probability function $P_d(p)$. In turn, this implies that an average of $\mu(p) := pP_d(p)$ packets are delivered per time slot. A remarkable property of RA networks is that despite the lack of coordination among users, it is possible to achieve a reasonable average number of packets delivered by selecting p so as to achieve $\mu := \max[\mu(p)]$. In e.g., the slotted Aloha protocol, $\mu = 0.36$ which means that about 1 packet is delivered every 3 time slots.

In the present chapter, we discuss user cooperation in *random access* (RA) channels by drawing from two different sources. On the one hand, we draw from well-established spread spectrum random access (SSRA) protocols; see e.g., [2,43,62] and references therein. And on the other hand, we draw from the observation that user cooperation can be viewed as a form of multipath, a type of diversity for which SS with long PN sequences used as spreading codes is particularly well suited [76].

An intuitive notion underlying our contribution is that user cooperation is a form of diversity well matched to the very nature of RA networks. Indeed, the random nature of

RA dictates that at any given time only a fraction of potential users is active, the others having either empty queues or their transmissions deferred. Accordingly, given that only a few out of the total number of transmitters are active at any given time, transmission hardware resources are inherently under-utilized in wireless RA networks. As we will show, user cooperation can exploit these resources to gain a diversity advantage, without draining additional energy from the network and without bandwidth expansion. Reinforcing this intuitively reasonable notion, the number of temporarily idle users increases with the size of the network, indicating that user cooperation is available when most needed; i.e., in congested heavily-populated networks. While intuitive notions not always turn out to be true, this one will; the main purpose of this chapter being precisely to establish that as the network size increases, there is an increasing diversity advantage to be exploited leading to a limiting scenario in which *the throughput of cooperative RA over wireless fading channels approaches that of an equivalent system operating over an additive white Gaussian noise (AWGN) channel*.

Building on an existing network diversity multi-access (NDMA) protocol [109], cooperative RA has been also considered in [54, 55], where re-transmitting cooperators aid the separation of multiple collided packets. However, NDMA-based schemes are known to be challenged by channel ill-conditioning, difficulty in determining the number of collided packets and relatively high complexity at the access point as well as at the relays, which require analog (waveform storage and) forwarding [54, 55].

The rest of the chapter is organized as follows. The spatial distribution of users and the physical propagation model are introduced in Section 5.1 to formalize the notion that cooperation takes place among nearby users. In Section 5.1.1, we provide a high level description of how our cooperative RA protocol operates and explain different user states that emerge due to cooperation. We then introduce in Section 5.2 a novel non-cooperative SSRA protocol upon which a cooperative version is built later on. The throughput of this protocol is analyzed in Section 5.2.1 to serve as a benchmark as well as to illustrate the tools utilized. A consequence of this analysis, discussed in Section 5.2.2, is to motivate the beneficial role of diversity by showing how it can close the large throughput gap between

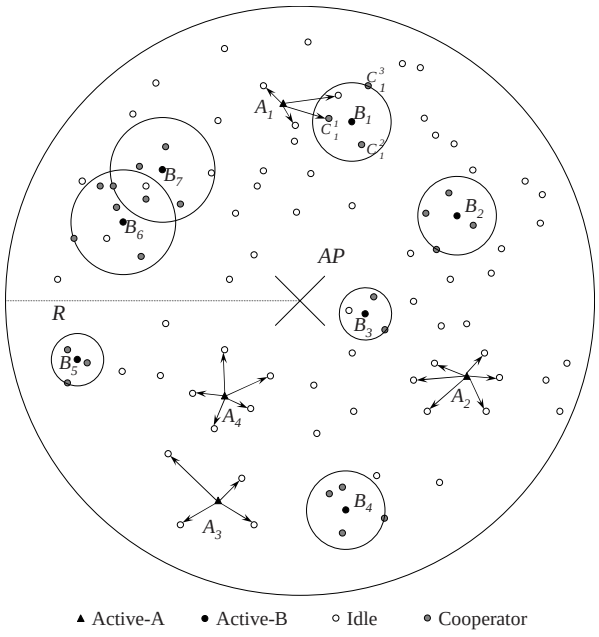

Figure 5.1: A snapshot of a cooperative RA network. Users are divided into four classes: Active-A users trying to reach nearby idle users, Active-B users trying to reach the AP, Idle users that have empty queues or deferred their transmissions, and Cooperators that are helping Active-B users in reaching the AP.

corresponding systems operating over wireless and wireline channels.

Having made the case for diversity, we argue about a symbiotic relation between RA and user cooperation and introduce in Section 5.3 our Opportunistic Cooperative Random Access (OCRA) protocol based on the opportunistic exploitation of highly reliable links among neighboring users. We then move on to study its throughput in Section 5.4 and introduce our main results regarding OCRA's asymptotic throughput as the number of users grows large in Section 5.4.1. Section 5.4.1 contains only the most relevant results, with a more detailed asymptotic behavior analysis postponed to Section 5.5, where we show how pertinent theorems formalize intuitive comments made in this introduction about the suitability of user cooperation as *the* form of diversity for RA networks. Finally, synchronization issues motivate an unslotted counterpart of OCRA that we present in Section 5.6.

5.1 Preliminaries

The problem addressed in this chapter is that of designing a cooperative RA protocol. Consider a set of J users, $\mathcal{J} = \{U_j\}_{j=1}^{J}$, communicating with an access point (AP) in a wireless RA network as depicted in Fig. 5.1. User j and its position in a coordinate system centered at the AP will be denoted by U_j. With these positions considered random and uniformly distributed within a circle of radius R, we express the probability of U_j to have distance from the AP smaller than r as

$$\Pr\{\|U_j\| < r\} = \frac{r^2}{R^2}, \qquad 0 \leq r \leq R, \tag{5.1}$$

where $\|U_j\|$ denotes the 2-norm of the position vector U_j. User positions are further assumed independent.

Users transmit blocks of duration T with U_j's block denoted as $\mathbf{x}_{U_j} := \{x_{U_j}(t)\}_{t=0}^{T-1}$. The broadcast nature of the wireless channel dictates that the signal $\mathbf{z}_{U_{j_1}} = \{z_{U_{j_1}}(t)\}_{t=0}^{T-1}$ received at any point is the superposition of all users' signals, $\{\mathbf{x}_{U_{j_2}}\}_{j_2=1}^{J}$; i.e.,

$$\mathbf{z}_{U_{j_1}} = \sum_{j_2=1}^{J} h(U_{j_2}, U_{j_1})\mathbf{x}_{U_{j_2}} + \mathbf{n}, \tag{5.2}$$

where $\mathbf{n} := \{n(t)\}_{t=0}^{T-1}$ is zero-mean additive white Gaussian noise (AWGN) with variance $\mathrm{E}[n^2(t)] = N_0$, and $h(U_{j_2}, U_{j_1})$ denotes the Rayleigh block fading channel coefficient corresponding to the link $U_{j_2} \to U_{j_1}$. When $U_{j_1} \equiv AP$ we will denote $z_{AP}(t) \equiv z(t)$ and $h(U_{j_2}, AP) \equiv h(U_{j_2})$.

The average power received at U_{j_1} from a source U_{j_2} transmitting with power $P(U_{j_2})$ adheres to an exponential path loss model

$$P(U_{j_2} \to U_{j_1}) = \frac{\xi\ P(U_{j_2})}{\|U_{j_1} - U_{j_2}\|^{\alpha}}, \tag{5.3}$$

with ξ and $\alpha \geq 2$ denoting the pathloss constant and exponent respectively [69, Chap.14]. As a special case, the power received at the AP from U_{j_2} is $P(U_{j_2} \to AP) = \xi P(U_{j_2})/\|U_{j_2}\|^{\alpha}$. Consistent with (5.3), the Rayleigh block fading coefficient $h(U_{j_1}, U_{j_2})$ in (5.2) is complex

Gaussian distributed with zero-mean and variance

$$\begin{aligned}\mathrm{var}[h(U_{j_1},U_{j_2})] &:= \mathrm{E}\left[h(U_{j_1},U_{j_2})h^*(U_{j_1},U_{j_2})\right] \\ &= \mathrm{E}\left[|h(U_{j_1},U_{j_2})|^2\right] \\ &= \frac{\xi}{\|U_{j_1}-U_{j_2}\|^{\alpha}}. \end{aligned} \tag{5.4}$$

We assume that fading coefficients linking different users are uncorrelated and that channel state information is obtained by the receivers (e.g., using a training sequence) to permit coherent reception. We further note that block fading coefficients $h(U_{j_1},U_{j_2})$ are constant for the duration of a transmission block but different and uncorrelated across blocks.

5.1.1 Two-phase cooperation

Transmission in the proposed cooperative RA protocol proceeds in two phases. In the first phase, "phase-A", the user sends a packet with sufficient power to be correctly decoded by nearby peers; while in the second phase, "phase-B", the set of peers that successfully decoded this packet transmit cooperatively with power sufficient to reach the AP. If we manage to balance conflicting power requirements, what will happen in phase-A is that nearby users decode the original packet while the power received at the destination is negligible. On the one hand, this implies that phase-A users do not interfere severely with nodes which are at the same time operating in phase-B. On the other hand, phase-A succeeds in locally disseminating information so that subsequent phase-B transmissions are enriched with a certain degree of user cooperation diversity.

It is not necessary to follow a given user from phase-A to phase-B, because what will happen to current phase-A users when they reach phase-B is statistically indistinguishable from what is happening to current phase-B users. It thus suffices to study a snapshot of the RA network which comprises current phase-A and phase-B users. At this given snapshot, the set of users $\mathcal{J}$ is temporarily divided into a set of N_A "active-A" users, $\mathcal{A}=\{A_j\}_{j=1}^{N_A}$, operating in phase-A of their transmission trying to reach nearby users; a set of N_B active-B users, $\mathcal{B}=\{B_j\}_{j=1}^{N_B}$, communicating their packets to the AP; and N_I idle users $\mathcal{I}=\{I_j\}_{j=1}^{N_I}$ that either have empty queues or decided not to transmit. Clearly,

we have that $\mathcal{J} = \mathcal{A} \cup \mathcal{B} \cup \mathcal{I}$. A fourth class of users, encompasses the sets of cooperators $\mathcal{C}_j = \{C_j^k\}_{k=0}^{K_j}$ associated with each active-B user B_j. The set $\mathcal{C}_j$ contains the K_j users that correctly decoded B_j's phase-A packet in the previous slot, and we adopt the convention that $C_j^0 = B_j$.

Remark 13 It is worth stressing that the different sets of users are not necessarily mutually exclusive. Actually, the sole constraint on the classes is

$$\mathcal{I} \cap (\mathcal{B} \cup \mathcal{A}) = \emptyset, \tag{5.5}$$

meaning that a terminal cannot be idle and active-A or active-B at the same time, but is allowed to be active-A and active-B in the same slot, as we will detail later. Also, it is convenient to regard cooperators as a parallel class in the sense that

$$\mathcal{C}_j \subseteq \mathcal{I} \cup \mathcal{A}, \tag{5.6}$$

implying that a cooperator is either regarded as active-A, if it independently decided to transmit its own information, or as idle, if it did not. The reason for these requirements will become clear in Section 5.3.

It will turn out, that phase-A will be the phase determining the system's performance; a perhaps intuitive result since it is in this phase that the need arises to balance the conflicting requirements of transmitting with as low power as possible while reaching as many idle users as possible. To this end, we will isolate one of the statistically identical phase-A user nodes, call it $U_0 \in \mathcal{A}$, and study the tradeoff between phase-A power and number of idle users reached. Without loss of generality, we further assume that $U_0 = A_{N_A}$. Let $\mathcal{C}_0 = \{C_0^k\}_{k=0}^{K_0}$ denote the set of (idle) users that successfully decode U_0's phase-A packet with the convention that $C_0^0 = U_0$. Note that the nodes in the set $\mathcal{C}_0$ are not cooperating with U_0 in the current slot, but will do so in the next one.

The key to delineate the aforementioned power tradeoff is to observe that the closer an idle node is to U_0 the larger is the probability of decoding U_0's active-A packet correctly. Consequently, we will consider distance-ordered sets with $I_0^{(k)}$, $A_0^{(k)}$ and $B_0^{(k)}$ denoting the

k^{th} closest to U_0, idle, active-A and active-B user respectively[1]; i.e.,

$$\|I_0^{(1)} - U_0\| \leq \|I_0^{(2)} - U_0\| \leq \ldots \leq \|I_0^{(k)} - U_0\|, \qquad I_0^{(1)} \ldots I_0^{(k)} \in \mathcal{I} \ , \tag{5.7}$$

with similar expressions holding true for active-A and active-B users. Likewise, we will order the sets of cooperators according to their distance from the active-B user they are cooperating with

$$0{=}\|C_j^{(0)}{-}B_j\| \leq \|C_j^{(1)}{-}B_j\| \leq \ldots \leq \|C_j^{(K_j)}{-}B_j\|, \ j \in [1, N_B] \tag{5.8}$$

where the first equality follows from the convention $C_j^{(0)} = B_j$.

Note that consistent with the random nature of RA networks, the degree of cooperation K_j that each U_j receives is itself random, not requiring pre-established agreement among users. Cooperative RA throughput will be determined by the statistics of K_j, the characterization of which constitutes a central topic of this chapter.

5.2 Non-Cooperative SS Random Access

In this section, we present a non-cooperative spread spectrum (SS) RA protocol upon which we will build the cooperative version in Section 5.3. While many such non-cooperative SSRA systems have been proposed and analyzed in the literature (see e.g., [2,43,62] and references therein) we summarize here the one introduced in [127] that we regard as the best starting point for our cooperative protocol in Section 5.3. The queue model is depicted in Fig. 5.2, where each of the J users has an infinite-length buffer for storing L-bit fixed length packets that arrive at a rate of λ packets per packet duration. The packet arrival processes are identically distributed (i.d.), not necessarily independent, yielding a total arrival rate of $J\lambda$ packets per packet duration.

The L bits of each packet are spread by a factor S (a.k.a. spreading gain) to construct a transmitted packet of $T := SL$ chips. Spreading is implemented using a long pseudo-noise (PN) sequence $\mathbf{c} := \{c(t)\}_{t\in\mathbb{Z}}$ with period $\mathcal{P} = SL = T$. Letting $\mathbf{d}_{U_j} := \{d_{U_j}(l)\}_{l=0}^{L-1}$ denote

[1]Subscripts and superscripts in parentheses will henceforth signify ordering.

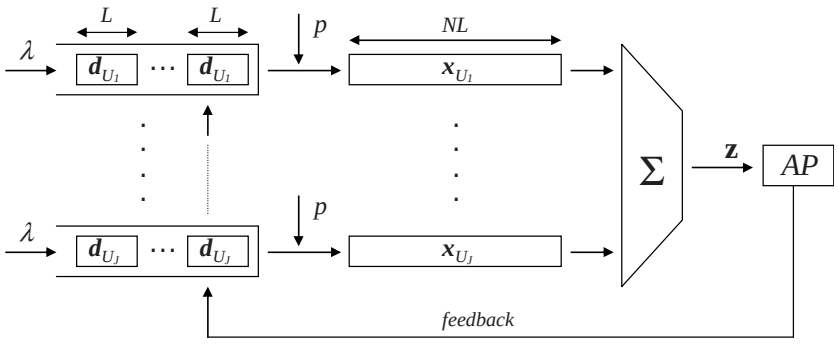

Figure 5.2: Queue and transmission diagram of a non-cooperative SSRA network. Packets are spread using random shifts of a common long PN sequence.

a *data* packet of user U_j, and $\mathbf{x}_{U_j} := \{x_{U_j}(t)\}_{t=0}^{T-1}$ the corresponding *transmitted* packet, we have

$$x_{U_j}(Sl+s) = \sqrt{P(U_j)}\; d_{U_j}(l)c(Sl+s-\tau_{U_j}), \qquad (5.9)$$
$$l \in [0, L-1],\ s \in [0, S-1],$$

where we note that $\mathbf{c}$ is a common long PN sequence *shared* by all users, τ_{U_j} is a user-specific shift applied to $\mathbf{c}$, and $P(U_j)$ is the power transmitted by node U_j.

These spread packets are transmitted to the AP, which acknowledges successfully decoded packets through a common feedback channel. As in [2,43,62] feedback is assumed to be instantaneous and free of errors.

We are now ready to define the non-cooperative SSRA protocol considered in this chapter by the following rules:

[**R1**] Time is divided into slots, each comprising T chip periods. If users decide to transmit, they do so at the beginning of a slot.

[**R2**] Packets are spread for transmission according to (5.9). The shift τ_{U_j} is selected at random by each user; and $P(U_j) = P_0\|U_j\|^\alpha/\xi$ effects average power control so that all users are received at the AP with the same average power P_0 [c.f. (5.3)].

[**R3**] If a given user's queue is not empty, the user transmits the first queued packet in the next slot with probability p.

Rule [R1] defines a slotted system and its purpose is to simplify throughput analysis; [R2] effects statistical user separation and power control; and [R3] controls the transmission rate, with p adjusted so as to maximize throughput.

To better appreciate [R2], let $N \leq J$ denote the number of users active in a given slot and consider the block $\mathbf{z} := \{z(t)\}_{t=0}^{T-1}$ received at the AP. Specializing (5.2) to the superposition of these N transmissions, the received chips (entries of $\mathbf{z}$) are

$$z(Sl+s)=\sum_{j=1}^{N}\sqrt{P(U_j)}h(U_j)d_{U_j}(l)c(Sl+s-\tau_{U_j})+n(Sl+s). \tag{5.10}$$

To recover packets from a given user, say U_N without loss of generality, we compensate the random phase by multiplying with the normalized channel conjugate $h_n^*(U_N) := h^*(U_N)/|h(U_N)|$ and despread $\mathbf{z}$ using the properly delayed version of the long PN sequence $\mathbf{c}(t-\tau_{U_N})$. The resultant decision vector $\mathbf{r}_{U_N} := \{r_{U_N}(l)\}_{l=0}^{L-1}$ has entries

$$\begin{aligned} r_{U_N}(l) &= \frac{h_n^*(U_N)}{S}\sum_{s=0}^{S-1} z(Sl+s)c(Sl+s-\tau_{U_N}) \\ &= \sqrt{P(U_N)}|h(U_N)|d_{U_N}(l)+\sum_{j=1}^{N-1}\mathbb{I}(l;U_N\rightarrow AP;U_j)+\tilde{n}(l) \end{aligned} \tag{5.11}$$

where we used $h(U_N)h_n^*(U_N) \equiv |h(U_N)|$. Note that the noise variance is reduced by S; i.e., $\text{var}[\tilde{n}(l)] = N_0/S$, and interference terms emerge due to users $\{U_j\}_{j=1}^{N-1}$; the symbol $\mathbb{I}(l;U_{j_0}\rightarrow AP;U_j)$ denotes the interference of user U_j to the communication of bit l from U_N to the AP, and is given by

$$\begin{aligned} \mathbb{I}(l;U_N\rightarrow AP;U_j) =& \frac{1}{S}\sqrt{P(U_j)}\; h(U_j)h_n^*(U_N)d_{U_j}(l) \\ & \times \sum_{s=0}^{S-1} c(Sl+s-\tau_{U_j})c(Sl+s-\tau_{U_N}). \end{aligned} \tag{5.12}$$

The most important property of PN sequences is that they have a white-noise like autocorrelation $\text{E}[c(t-\tau_{U_j})c(t-\tau_{U_N})] \approx \delta(\tau_{U_j}-\tau_{U_N})$, from where we deduce that if $\tau_{U_j} \neq \tau_{U_N}$, then

$$\text{E}[\mathbb{I}(l;U_N\rightarrow AP;U_j)] = 0 \tag{5.13}$$

$$\text{var}[\mathbb{I}(l;U_N\rightarrow AP;U_j)] = P_0/S \tag{5.14}$$

$$\text{E}[\mathbb{I}(l;U_N\rightarrow AP;U_{j_1})\mathbb{I}^*(l;U_N\rightarrow AP;U_{j_2})] = 0,\ \forall j_1 \neq j_2 \tag{5.15}$$

where in deriving the last equality we also exploited the independence of users' fading coefficients when $\tau_{U_{j_1}} = \tau_{U_{j_2}}$.

Combining (5.11) with (5.13), we find readily that the expected value of the decision vector is

$$\mathrm{E}[r_{U_N}(l)] = \sqrt{P(U_N)}|h(U_N)|d_{U_N}(l) \quad l \in [0, L-1], \tag{5.16}$$

from where it follows that a suitable demodulator is $\hat{\mathbf{d}}_{U_{j_0}} = \mathrm{sign}(\mathbf{r}_{U_{j_0}})$. The interference increases the variance of the decision variable $r_{U_N}(l)$ in (5.11), which after using properties (5.14) and (5.15) turns out to be

$$\mathrm{var}[r_{U_N}(l)] = N_0/S + P_0(N-1)/S, \quad l \in [0, L-1]. \tag{5.17}$$

Eq. (5.17) implies that the interference increases the probability of error because it increases the variance of the decision statistic. As in e.g., [114, Chap.2] we can model the interference as Gaussian and independent for different bits, implying that the probability that a packet is correctly decoded is fully determined by the signal to interference-plus-noise ratio (SINR). When N users are active, the instantaneous SINR is [c.f. (5.16) and (5.17)]

$$\gamma_N := \frac{\mathrm{E}^2[r_{U_N}(l)]}{\mathrm{var}[r_{U_N}(l)]} = S\frac{P(U_N)|h(U_N)|^2}{N_0 + (N-1)P_0}, \tag{5.18}$$

and the average SINR is found by taking expected values with respect to the channel distribution [c.f. (5.18)]

$$\bar{\gamma}_N := \mathrm{E}[\gamma_N] = \frac{S}{N_0/P_0 + N - 1}, \tag{5.19}$$

where we used that $P(U_N)|\mathrm{E}[h(U_N)|^2] = P_0$ which follows from the average power control in [R2] and the channel model in (5.3).

We established in (5.16) that through [R2] we effect statistical separation of different users' packets, with packet error probability (PEP) determined by the SINR in (5.19). Notice though, that there is also a chance to have $\tau_{U_j} = \tau_{U_N}$ for some $j \neq N$. Both this and the interference term will determine the throughput of this non-cooperative RA protocol, motivating a distinction between what we term soft and hard collisions which we define as follows.

Definition 5 *(Soft and hard collisions)*

[a] *We say that U_{j_0} experiences a "hard collision" (HC) if $\tau_{U_{j_0}} = \tau_{U_j}$ for some $j \neq j_0$; the HC event is*

$$\mathrm{HC}(U_{j_0}) := \bigcup_{j \neq j_0} \left\{ \tau_{U_{j_0}} = \tau_{U_j} \right\}. \tag{5.20}$$

[b] *Given that U_{j_0} does not experience a hard collision, we say that it experiences a "soft collision" (SC) when the packet is lost due to interference:*

$$\mathrm{SC}(U_{j_0}) := \{ \hat{\mathbf{d}}_{U_{j_0}} \neq \mathbf{d}_{U_{j_0}} \mid \mathrm{HC}^c(U_{j_0}) \}, \tag{5.21}$$

where $\mathrm{HC}^c(U_{j_0})$ denotes the complement of $\mathrm{HC}(U_{j_0})$.

Conditioned on the number of active users N, we can evaluate the probability that U_{j_0} experiences a HC as the probability that any of the $N-1$ interferers chooses the same PN shift

$$\begin{aligned} P_{\mathrm{HC}}(N) :=& \Pr\{\mathrm{HC}(U_{j_0}) \mid N\} = 1 - \Pr\{\mathrm{HC}^c(U_{j_0}) \mid N\} \\ &= 1 - \left(1 - \frac{1}{T}\right)^{N-1}, \end{aligned} \tag{5.22}$$

where we used that since there are T possible PN shifts, $\Pr\{\tau_{U_{j_0}} = \tau_{U_j}\} = 1/T$. Likewise, the SC probability $P_{\mathrm{SC}}(N)$ can be inferred from the SINRs in (5.18) and (5.19). For a given channel realization $h(U_N)$, $P_{\mathrm{SC}}(N)$ is a function of the instantaneous SINR in (5.18); however, what matters from a throughput perspective is $P_{\mathrm{SC}}(N)$ averaged over all channel realizations. We thus write

$$P_{\mathrm{SC}}(N) := \Pr\{\mathrm{SC} \mid N\} = P_e\left(\bar{\gamma}_N\right)\left[1 - P_{\mathrm{HC}}(N)\right], \tag{5.23}$$

where $P_e(\bar{\gamma}_N)$ is a function that maps the link *average* SINR, $\bar{\gamma}_N$, to the *average* PEP. The function $P_e(\bar{\gamma}_N)$ is determined by the channel model and the transmission/reception parameters which include the type of modulation, type of receiver and forward error correcting (FEC) code. The existence of $P_e(\bar{\gamma}_N)$ is guaranteed since we model the interference as Gaussian and independent across bits. In fact, given Rayleigh interferers $P_e(\bar{\gamma}_N)$ is also a function of N, S and P_0/N_0 as clarified in Remark 15.

A packet is successfully decoded if and only if it neither experiences a hard collision, nor a soft one. Accordingly, the packet success probability with N active users ($N-1$ interferers) is

$$P_s(N) := 1 - P_{\mathrm{HC}} - P_{\mathrm{SC}} = \left(1 - \frac{1}{T}\right)^{N-1} [1 - P_e(\bar{\gamma}_N)]. \tag{5.24}$$

The throughput of this non-cooperative SSRA system can be obtained from (5.24) as we analyze in the next section.

5.2.1 Throughput Analysis

A possible performance measure of RA networks is the average departure rate μ; if we let $P_s = \sum_{n=1}^{J} \Pr\{N = n\} P_s(n)$ be the probability that a packet transmitted by the reference user U_{j_0} is successfully decoded by the AP, then

$$\mu = pP_s. \tag{5.25}$$

However, throughput instead of departure rate is the standard metric whose definition follows from the concept of stability. We let $\mathtt{q}_j(m)$ be the number of packets in U_j's queue in the m^{th} slot, and say that this queue is stable if, [58]

$$\lim_{m\to\infty} \Pr\{\mathtt{q}_j(m) \le x\} = \mathtt{Q}(x) \quad \text{with} \quad \lim_{x\to\infty} \mathtt{Q}(x) = 1. \tag{5.26}$$

The conditions in (5.26) assert that the system is stable if and only if there exists a positive probability mass function of $\{\mathtt{q}_j(m)\}_{j=1}^{J}$ when $m \to \infty$. A system is called stable if all the queues are stable, and throughput is defined as follows:

Definition 6 *The maximum aggregate throughput is defined as the unique quantity η such that the system is stable if $J\lambda < \eta$ and unstable if $J\lambda > \eta$.*

Thus, η is defined as the maximum aggregate arrival rate that the system can afford with stable queues. If $J\lambda < \eta$, then individual queues have a bounded number of packets and the packets get transmitted with finite delay. If $J\lambda > \eta$, then the queues grow without limit and the packets experience infinite delays.

The system will be clearly unstable if $\lambda > \mu$. Accordingly, the throughput cannot exceed the departure rate $\eta \leq J\mu$. What is not so obvious is whether $\lambda < \mu$ yields a stable system. Indeed, this is not true in general but for symmetric and stationary systems it is true due to Loynes' theorem [58]. For this subclass of systems, we thus have

$$\eta = J\mu. \tag{5.27}$$

A challenge with the protocol defined by rules [R1]-[R3] is that the service processes are not necessarily stationary due to the possibility of having empty queues. Notwithstanding, by resorting to a dominant system approach, [111], and following an equivalence argument (see [23, 71]), we can establish that $\eta = J\mu$ for the SSRA protocol introduced in Section 5.2 to obtain the following proposition.

Proposition 4 *Consider the protocol defined by rules* [R1]-[R3]*, and not necessarily independent but i.d. arrival processes with rate* λ*. Then, the average aggregate throughput is*

$$\begin{aligned} \eta =& \eta(J, N_0/P_0, S, p) \\ :=& Jp \sum_{n=0}^{J-1} \binom{J-1}{n} p^n (1-p)^{J-1-n} \left(1 - \frac{1}{T}\right)^n [1 - P_e(\bar{\gamma}_{n+1})] \end{aligned} \tag{5.28}$$

with $\bar{\gamma}_{n+1} := 1/(N_0/P_0 + n/S)$.

Proof: Define the dominant system by replacing rule [R3] with:

[R3'] Users transmit with probability p. If a user's queue is empty, then the corresponding user transmits a dummy packet.

Rule [R3'] is commonly used to decouple the different users' queues. But here we are interested in the fact that it renders the system stationary and allows application of Loynes' theorem. Thus, using (5.27) for the dominant system we have

$$\eta_{DS} = J\mu = JpP_s, \tag{5.29}$$

with η_{DS} denoting the dominant system's throughput.

To compute P_s, we condition on the number of *interfering* users $N-1$ to obtain

$$\begin{aligned} P_s &= \sum_{n=0}^{J-1} \Pr\{N-1=n\} P_s(n+1) \\ &= \sum_{n=0}^{J-1} \Pr\{N-1=n\} \left(1-\frac{1}{T}\right)^n [1-P_e(\bar{\gamma}_{n+1})], \end{aligned} \tag{5.30}$$

where the limits of the summation are because the number of interferers is between 0 and $J-1$, and the second equality follows from (5.24) with $N-1=n$.

On the other hand, since interferers act independently $N-1$ follows a binomial distribution with parameters p and $J-1$ and accordingly $\Pr\{N-1=n\} = \binom{J-1}{n} p^n (1-p)^{J-1-n}$, which upon substitution into (5.30) yields

$$P_s = \sum_{n=0}^{J-1} \binom{J-1}{n} p^n (1-p)^{J-1-n} \left(1-\frac{1}{T}\right)^n [1-P_e(\bar{\gamma}_{n+1})]. \tag{5.31}$$

Furthermore, substituting (5.31) into (5.29) yields (5.28) and establishes the result for the dominant system defined by rules [R1], [R2] and [R3'].

We can now repeat the argument in [23], for what we consider identical instantiations of the arrival processes fed to the dominant and original systems. Given that we are adding (dummy) packets, the queues in the fictitious dominant system can never be shorter than the queues in the original system. It follows that if the dominant system is stable, then so must be the original system; hence $\eta \geq \eta_{DS}$. Assume now that $\eta > \eta_{DS}$ strictly, to infer that there exists an arrival rate $\eta > \lambda J > \eta_{DS}$ that makes the original system stable and the dominant system unstable. But this is a contradiction since if the dominant system were unstable, there would be no long-term need for dummy packets since all the queues in the dominant system would eventually become continuously backlogged with real packets. The dominant system is therefore equivalent to the original system; hence, the original system is also unstable. So, we must have $\eta = \eta_{DS}$, and (5.28) is also valid for the original system defined by rules [R1]-[R3]. □

Note that η in (5.28) is a function of the number of users J, the noise to signal ratio N_0/P_0, the spreading gain S and the transmission probability p. We are usually interested

in the maximum stable throughput (MST) defined as

$$\eta_{\max}(J, N_0/P_0, S) = \max_p \{\eta(J, N_0/P_0, S, p)\}, \tag{5.32}$$

and achieved at $p = p_{\max}$. In this particular work, we will be interested in the asymptotic MST that we define as

$$\eta_\infty(N_0/P_0, S) = \lim_{J\to\infty} \eta_{\max}(J, N_0/P_0, S), \tag{5.33}$$

and interpret as the average number of packets transmitted per unit time in a system with a very large number of users.

In Section 5.4, we will compare η_∞ for the SSRA protocol introduced in this section against a suitably defined cooperative RA protocol. Before moving on to that, let us show what advantage diversity has to offer in RA systems.

5.2.2 On the role of diversity in RA

For this section only, we consider different models for the channels $h(U_j)$ and present a motivating example of the function $P_e(\bar{\gamma}_N)$. Let us suppose that we use BPSK modulation with coherent detection and code the packet with a BCH block code capable of correcting up to $\epsilon_{\max}$ errors. With $Q(x) := (1/\sqrt{2\pi}) \int_x^\infty e^{-u^2/2}\, du$ denoting the Gaussian tail function and recalling the Gaussian model of interference, the bit error probability with γ_N instantaneous SINR is $q(\gamma_N) = Q(\sqrt{2\gamma_N})$ [69, sec. 5.2] and the corresponding *instantaneous* PEP is given by [69, p.437]

$$P_{e,i}(\gamma_N) = 1 - \sum_{\epsilon=0}^{\epsilon_{\max}} \binom{L}{\epsilon} q^\epsilon(\gamma_N)[1 - q(\gamma_N)]^{L-\epsilon}. \tag{5.34}$$

It is interesting to compare the throughput as determined by (5.28) for different channel models. The best possible scenario is when $h(U_j)$ is a deterministic constant (AWGN channel), in which case $\gamma_N = \bar{\gamma}_N$ and the corresponding average PEP is thus $P_e^G(\bar{\gamma}_N) = P_{e,i}(\gamma_N)$.

A better model for the wireless environment however, is a Rayleigh fading channel where γ_N is random Rayleigh distributed (since $|h(U_j)|^2$ is). In this case, we have to average (5.34)

over the channel (Rayleigh) distribution $f_{\gamma_N}(\gamma_N)$ to obtain

$$P_e^R(\bar{\gamma}_N) = \int_0^\infty P_e(\gamma_N) f_{\gamma_N}(\gamma_N). \tag{5.35}$$

It can be easily verified that for moderate and large $\bar{\gamma}_N$ we have $P_e^R(\bar{\gamma}_N) \gg P_e^G(\bar{\gamma}_N)$, ultimately leading to a much smaller throughput when otherwise equivalent systems operate over Rayleigh channels than when they operate over AWGN channels.

The throughput over wireless channels can be increased with diversity techniques, e.g., multiple transmit antennas. Consider a terminal with κ antennas transmitting a packet as in (5.9) using a user and antenna-specific $\tau_{U_j,\kappa}$ so that despreading $\mathbf{z}$ in (5.10) with $\mathbf{c}(t - \tau_{U_j,\kappa})$ recovers the signal transmitted by U_j's κ^{th} antenna. This way the AP can decode κ copies received through uncorrelated Rayleigh channels, $\{h_k(U_j)\}_{k=1}^{\kappa}$, yielding the aggregate channel model $|h(U_j)|^2 := \sum_{k=1}^{\kappa} |h_k(U_j)|^2$ when maximum ratio combining is used. If we let the uncorrelated channels have equal average received powers so that $P(U_j)\mathrm{E}[|h_k(U_j)|^2] = P_0/\kappa$, the channel distribution $f_{\gamma_N}(\gamma_N)$ is chi-square with 2κ degrees of freedom. To fully characterize this distribution we repeat steps (5.11) - (5.19) to obtain the per-path average SINR

$$\bar{\gamma}(N,\kappa) := S\frac{1/\kappa}{N_0/P_0 + N - 1 + (\kappa-1)/\kappa}, \tag{5.36}$$

where in the denominator, the term N_0/P_0 comes from the AWGN, the term $N-1$ from the interference from other terminals and the term $(\kappa-1)/\kappa$ from the (self-)interference of the remaining $\kappa-1$ paths of the same terminal. The corresponding aggregate SINR is given by $\bar{\gamma}_N := \kappa\bar{\gamma}(N,\kappa)$ and the average PEP $P_e^{\kappa}(\bar{\gamma}_N)$ can be found from (5.35) with $f_{\gamma_N}(\gamma_N)$ modified accordingly [69, sec. 14.4].

A particularly important fact for the present work is that if $\kappa \to \infty$ in the κ-order diversity channel, then the channel $|h(U_j)|^2$ approaches an AWGN channel. Indeed,

$$\lim_{\kappa\to\infty} P(U_j)|h(U_j)|^2 = \lim_{\kappa\to\infty} \kappa\,\frac{1}{\kappa}\sum_{k=1}^{\kappa} P(U_j)|h_k(U_j)|^2 = P_0, \tag{5.37}$$

where the limit follows from $\mathrm{E}[|h_k(U_j)|^2] = P_0/\kappa$ and the strong law of large numbers. But (5.37) implies that $|h(U_j)|^2$ converges to a constant which by definition leads to an

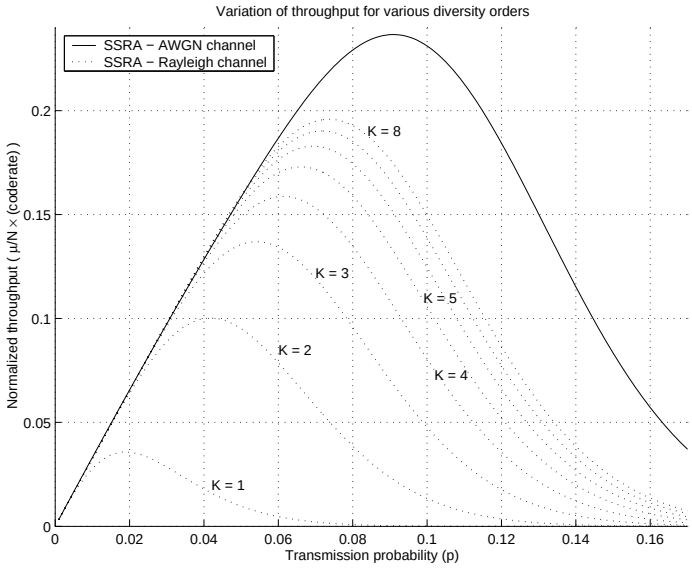

Figure 5.3: High-order diversity closes the enormous gap between the performance of RA over wireless Rayleigh fading channels with respect to wireline AWGN channels ($J = 128$, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

AWGN channel. We can now take the limit in (5.36) to obtain

$$\lim_{\kappa\to\infty} \kappa\bar{\gamma}(N,\kappa) = \frac{S}{N_0/P_0 + N} = \bar{\gamma}_{N+1}. \tag{5.38}$$

And combine (5.38) with (5.37) to claim that as the diversity order $\kappa \to \infty$, the PEP $P_e^{\infty}(\bar{\gamma}_N) := \lim_{\kappa\to\infty} P_e^{\kappa}(\bar{\gamma}_N)$ of this ∞-order diversity channel approaches the PEP of a Gaussian channel with a (in most cases small) increase in SINR; i.e., $P_e^{\infty}(\bar{\gamma}_N) = P_e^{G}(\bar{\gamma}_{N+1})$.

For each of the channels considered, we depict in Fig. 5.3 the normalized throughput as a function of the transmission probability p. It comes as no surprise that the MST over a wireless (Rayleigh) channel is miserable, being almost an order of magnitude smaller than the MST of the wireline AWGN channel. Corroborating the implications of (5.37), this sizeable gap can be closed by diversity techniques, as hinted by the twofold increase observed with 2-order diversity and the close-to-AWGN MST enabled with 8-order diversity. We summarize this important observation in the following remark.

Remark 14 For a given ECC, let $\eta^G(J, N_0/P_0, S, p)$ and $\eta^{\kappa}(J, N_0/P_0, S, p)$ be the

throughput over an AWGN channel and a κ-order diversity channel, respectively. Defining the throughput over an ∞-order diversity channel as $\eta^{\infty}(J, N_0/P_0, S, p) := \lim_{\kappa\to\infty} \eta^{\kappa}(J, N_0/P_0, S, p)$ we can write [c.f. (5.19), (5.37) and (5.38)]

$$\eta^{\infty}(J, N_0/P_0, S, p) = \eta^{G}(J, N_0/P_0 + 1, S, p). \tag{5.39}$$

This also implies the same relation between MSTs and asymptotic MSTs, $\eta_{\infty}^{\infty}(N_0/P_0, S) = \eta_{\infty}^{G}(N_0/P_0 + 1, S)$, a fact that we will exploit later on in pertinent remarks. With the SNR before spreading being $N_0/P_0 \gg 1$ for usual values of SNR and S, we deduce that (5.39) entails almost identical throughputs.

To characterize the diversity advantage in the ensuing sections without resorting to a specific transmission/reception scheme, we introduce the following definition.

Definition 7 In the family of PEP functions $\{P_e^{\kappa}(\bar{\gamma}_N)\}_{\kappa\in\mathbb{N}}$, $P_e^{\kappa}(\bar{\gamma}_N)$ represents the PEP for a κ-order diversity channel when the SINR is $\bar{\gamma}_N := \kappa\bar{\gamma}(N, \kappa)$ with $\bar{\gamma}(N, \kappa)$ as in (5.36). Specifically, $P_e^{\kappa}(\bar{\gamma}_N)$ maps the average SINR to the average PEP for terminals with κ transmit antennas so that the information bearing signal is transmitted over κ independent Rayleigh channels $\{h_k(U_j)\}_{k=1}^{\kappa}$ with equal powers $P(U_j)h_k(U_j) = P_0/\kappa$ via user and antenna-specific PN delays $\tau_{U_j,\kappa}$.

An example of the family $\{P_e^{\kappa}(\bar{\gamma}_N)\}_{\kappa\in\mathbb{N}}$ is the one generated by BCH codes and described by (5.34) - (5.36). While in deriving these equations we used the Gaussian model of interference this assumption is not strictly necessary for our claims as we discuss in the following remark.

Remark 15 In deriving (5.17) we modelled the interference plus noise term $\sum_{j=1}^{N-1} \mathbb{I}(l; U_N \to AP; U_j) + \tilde{n}(l)$ in (5.11) as a Gaussian random variable independent for different values of l. This approximation was later used in this section to derive the PEP expressions (5.34) and (5.35). The Gaussian model of interference is often accurate in practice; more generally (and perhaps more importantly) though, Proposition 4 as well as other results derived in the ensuing sections are true regardless of this assumption. Indeed, what is relevant for our results is the existence of the family $\{P_e^{\kappa}(\bar{\gamma}_N)\}_{\kappa\in\mathbb{N}}$ in Definition 7. Clearly, $P_e^{\kappa}(\bar{\gamma}_N)$ can be defined in terms of the exact correlation of $\sum_{j=1}^{N-1} \mathbb{I}(l; U_N \to AP; U_j) + \tilde{n}(l)$ for different values

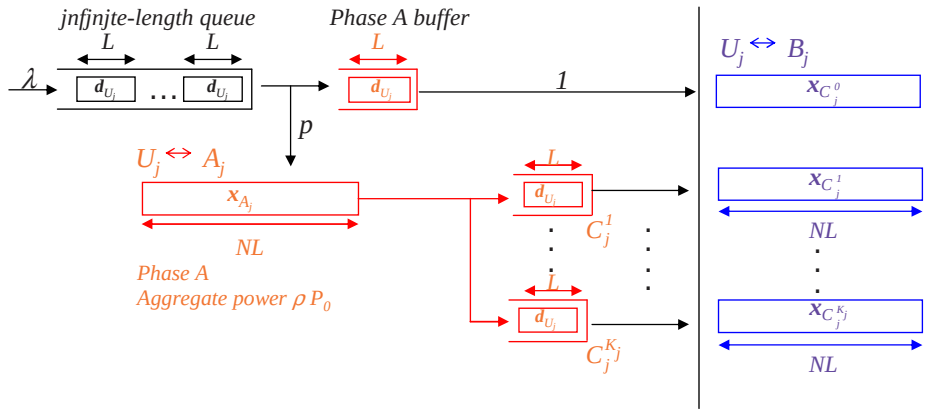

Figure 5.4: OCRA is a two-phase cooperative protocol. During phase-A users transmit with small power trying to recruit idle users as cooperators for phase-B. The seemingly conflicting requirements of small ρ and large K turn out to be asymptotically compatible.

of l. Note that if the interference plus noise is not modelled as independent for different bits l, P_e^{κ} depends on higher moments of the interference plus noise distribution and its characterization requires knowledge of N, S and Po/No. In our context of iid Rayleigh normalized channels, $P_e^{\kappa}(\bar{\gamma}_N)$ is in fact a function of only γ_N, N, S and Po/No. Since these three parameters are fixed throughout, we will write $P_e^{\kappa}(\bar{\gamma}_N)$ as a function of γ_N only and keep the rest implicit for brevity as in Definition 7. Also, even though the relation in (5.39) does not hold true without the Gaussian assumption, $P_e^{\infty}(\bar{\gamma}_N) = P_e^{G}(\bar{\gamma}_{N+1})$ still does. Moreover, as can be easily verified by simulations, $\eta^{\infty}(J, N_0/P_0, S, p) \approx \eta^{G}(J, N_0/P_0, S, p)$ as noted in Remark 14.

The present section has established that diversity offers the potential for a large throughput increase in RA networks; the point is, of course, whether and how this diversity can be enabled. This is the theme we deal with in the next section, where we explore the suitability of user cooperation to enable high order diversity in random access networks.

5.3 Opportunistic Cooperative Random Access

Because users transmit at random in RA networks a number of users remain idle over any given slot. Moreover, the transmission probability $p_{\max}$ that achieves MST decreases as J increases and the percentage of temporarily idle users that do not transmit in a given slot increases. This implies that a large number of potential cooperators (idle users) are available per active user and motivates user cooperation as a suitable diversity enabler for wireless RA.

Indeed, this large number of potential cooperators suggests a high probability of some of them having a good signal reception of any given user. The Opportunistic Cooperative Random Access (OCRA) protocol introduced in this section exploits this potential advantage since it relies on idle users with good reception opportunities. OCRA is a two-phase protocol as described in Section 5.1.1 and is defined by the following operating conditions; see also Fig. 5.4.

[S0] Let κ be a constant limiting the maximum achievable diversity. The period of the PN spreading code $\mathbf{c}(t)$ is chosen to be $\mathcal{P} = \kappa T + 1$.

[S1] At the beginning of each slot, if U_j's queue is not empty, U_j enters phase-A with probability p and moves the first packet in the queue, $\mathbf{d}_{U_j} := \{d_{U_j}(l)\}_{l=0}^{L-1}$, to a single-packet buffer that we term phase-A buffer.

[S2] **Phase-A**: When in phase-A, we say that $U_j \leftrightarrow A_j$ is an active-A user and transmits a packet spread according to (5.9) with PN-shift and power given by [c.f. [R2]]

$$\tau_{A_j} = 0, \qquad P(A_j) = \rho P_0 \|A_j\|^{\alpha}/\xi, \tag{5.40}$$

with $\rho \in (0, 1)$. Notice that the PN shift is deterministically chosen and the transmission power is so that the packet is received at the AP with fractional power ρP_0. A random integer, $\tau_{B_j} \sim \mathcal{U}[1, T]$, uniformly chosen over $[1, T]$ is included in the packet header to coordinate PN-shifts during phase-B. Let this transmitted packet be denoted as $\mathbf{x}_{A_j} := \{x_{A_j}(t)\}_{t=0}^{T-1}$.

[S3] Phase-A handshake: Any idle user I_k that successfully decodes $\mathbf{x}_{A_j}$ becomes a cooperator $I_k \leftrightarrow C_j^k$ and places $\mathbf{d}_{U_j}$ in a single-packet buffer designated for cooperation purposes. This successful decoding is acknowledged to A_j who collects a total of K_j acknowledgments and feeds forward the number K_j to the cooperators. Similar to e.g., [2, 43, 62] this handshake is assumed to be instantaneous and error free.

[S4] User U_j enters phase-B in the slot immediately after entering phase-A.

[S5] Phase-B: Let $\mathcal{C}_j = \{C_j^k\}_{k=0}^{K_j}$ be the set of cooperators as defined in Section 5.1.1 comprising $C_j^0 = B_j \leftrightarrow U_j$ and the K_j cooperators recruited in phase-A. Each of the C_j^k transmits the packet $\mathbf{d}_{U_j}$ spread according to (5.9) using

$$\tau_{C_j^k} = \tau_{B_j} + \tau_k T, \qquad P(C_j^k) = \frac{P_0}{K_j + 1} \|C_j^k\|^\alpha / \xi, \tag{5.41}$$

with τ_{B_j} the number received in phase-A's packet header, and the integer $\tau_k \sim \mathcal{U}[0, \kappa - 1]$. The power scaling is so that the total received power at the destination is P_0. Let $\mathbf{x}_{C_j^k} := \{x_{C_j^k}(t)\}_{t=0}^{T-1}$ denote these transmitted packets.

The number of cooperators K_j will be henceforth termed the *"cooperation order"* of B_j and the number κ_j of PN shifts chosen by at least one cooperator will be called the *"diversity order"* of B_j.

[S6] AP acknowledgement: If the superposition of phase-B packets corresponding to B_j is successfully decoded, the AP acknowledges this event through a feedback channel. If an acknowledgement is not received, the packet $\mathbf{d}_{B_j}$ is placed back in B_j's queue. The cooperators discard this packet in any event.

[S7] Idle operation: When not transmitting, $U_j \leftrightarrow I_j$ correlates the received signal with $\{c(t)\}_{t=0}^{T-1}$ to detect phase-A packets transmitted by other (nearby) users.

OCRA is a formal description of the two-phase protocol outlined in Section 5.1 based on the *opportunistic* exploitation of nearby users that happen to have a favorable signal reception of a given user. Phase-A is defined in rule [S2] by which U_j becomes the active-A user A_j and transmits $\mathbf{x}_{A_j}$ with low power so as to reach nearby users while not interfering

with the AP, this last situation requiring $\rho \ll 1$. Phase-B is defined by rule [S5] in which the packet is transmitted with κ_j-order diversity by $U_j \leftrightarrow B_j$ plus K_j cooperators corresponding to the K_j idle users that successfully decoded U_j's transmission during phase-A. Note that the opportunistic nature of the protocol manifests in the random diversity order κ_j which depends on the number K_j of cooperators recruited and the random selection of shifts τ_k used by these cooperators. Let us also recall that user devices are half-duplex and can decode a single packet per slot when not transmitting.

Rules [S1], [S4] and [S6] govern the transition between idle and active-A/B states. The transition from idle to active-A happens with probability p as per [S1]; after entering phase-A, the user proceeds deterministically to phase-B in the first upcoming slot ([S4]), and in most cases back to idle in the second one ([S6]). A lost packet does not alter this transition but only determines whether the packet is put back in queue or not. Also, [S6] dictates that cooperators do not keep track of acknowledgements discarding B_j's packet regardless of the transmission success. OCRA's complete transition diagram is slightly more involved due to the possibility of concurrent events. While most transitions are between idle and cooperator states and around the cycle idle to active-A to active-B to idle, other transitions and mixed states are also possible. Indeed, there is a chance for e.g., a user to be active-A and active-B in the same slot, or active-A and cooperator; also, instead of moving from active-B to idle we can move back to active-A if we independently choose to transmit a different packet. The complete transition diagram is shown in Fig. 5.5.

Rules [S0], [S3] and [S7] guarantee logical consistency of the protocol. According to [S0], the number of possible PN shifts is increased with respect to non-cooperative SSRA to enable the PN shift selection rule in phase-B [c.f. (5.41)]; [S3] disseminates the number of cooperators recruited to allow proper power scaling during phase-B as required by (5.41); and [S7] ensures that idle users are listening for phase-A packets.

A delicate issue in OCRA's description is the use of PN shifts, that is judiciously chosen to satisfy two requirements that we summarize in the following remark.

Remark 16 The PN shifts during phases A and B are selected in order to:

[a] Facilitate decoding of phase-A's packet by idle users. Indeed, since phase-A packets

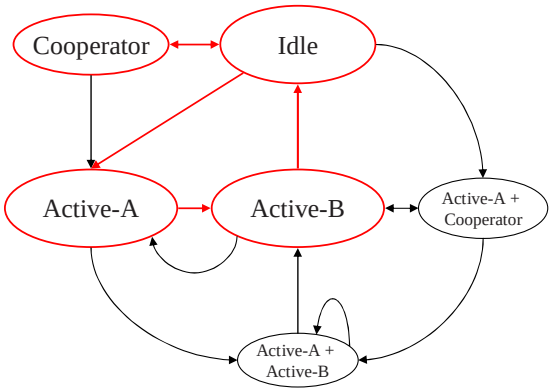

Figure 5.5: Most of the transitions are between Idle and Cooperator and from Idle to Active-A to Active-B and back to Idle. Some less common transitions are also possible.

use a fixed shift ($\tau_{A_j} = 0$), the idle users just need to correlate with a fixed sequence.

[b] Let the AP combine different cooperative copies of the same packet. If $\tau_{B_{j_1}} \neq \tau_{B_{j_2}}$, then $\tau_{C_{j_1}^{k_1}} \neq \tau_{C_{j_2}^{k_2}}$ $\forall k_1, k_2$, as can be seen from (5.41). Thus, if

$$\tau_{C_{j_1}^{k_1}} - \tau_{C_{j_2}^{k_2}} = \kappa_0 T \tag{5.42}$$

for some integer $\kappa_0 \in [0, \kappa - 1]$, then either the packets contains the same information, i.e., $j_1 = j_2$, or a hard collision occurred i.e. $\tau_{B_{j_1}} = \tau_{B_{j_2}}$.

Depending on their distances to the AP any user U_j experiences a propagation delay ω_{U_j}, so that if the latter is measured in chips, the PN shifts at the AP are perceived as $\tau_{U_j} + \omega_{U_j}$. While for SSRA propagation delays only add a random quantity ω_{U_j} to the already random τ_{U_j}, the remark in [b] is no longer valid for OCRA once we account for the propagation delay $\omega_{C_j^k}$. A simple solution used in e.g., the IS-95 standard [1], is to restrict the set of allowed shifts to a subset so that the difference in PN shifts is always larger than the maximum propagation delay, i.e., $\tau_{U_{j_1}} - \tau_{U_{j_2}} > \max_{[1,J]}\{\omega_{U_j}\}$.

Remark 16 is important in practice. A third consequence of the selection of PN shifts having theoretical as well as practical significance is given in the following proposition.

Proposition 5 *Given a slot with N_B active-B users, OCRA's hard collision probability (see Definition* 5-[a]*) for any reference user B_{j_1} is*

$$P_{\mathrm{HC}}(N_B) = 1 - \left(1 - \frac{1}{T}\right)^{N_B - 1}, \tag{5.43}$$

independently of the number of active-A users and cooperators' sets.

Proof: To evaluate this probability, note that $\tau_{C_{j_1}^{k_1}} = \tau_{C_{j_2}^{k_2}}$ can happen in two circumstances. The first is $j_2 = j_1$, in which case $\tau_{k_2} = \tau_{k_1}$ leads to $\tau_{C_{j_2}^{k_2}} = \tau_{C_{j_1}^{k_1}}$ according to (5.41). But in this case, both packets contain the same information and this is *not* a collision but just lost diversity[2] [c.f. Remark 16-[b]].

The second is $\tau_{B_{j_2}} = \tau_{B_{j_1}}$ for $j_2 \neq j_1$, in which case according to Remark 16-[b] the packets are combined as belonging to the same user. Thus, the hard collision event HC is equivalent to

$$\mathrm{HC} = \bigcup_{j_2 \neq j_1} \left\{ \tau_{B_{j_1}} = \tau_{B_{j_2}} \right\}. \tag{5.44}$$

Taking probabilities in (5.44) yields the expression

$$P_{\mathrm{HC}}(N_B) = 1 - \prod_{\substack{j_2=1 \\ j_2 \neq j_1}}^{N_B} \Pr\left\{ \tau_{B_{j_1}} \neq \tau_{B_{j_2}} \right\}. \tag{5.45}$$

But since the shifts $\tau_{B_{j_2}}$ are chosen uniformly and independently in $[1, T]$, we find that $\Pr\left\{ \tau_{B_{j_1}} \neq \tau_{B_{j_2}} \right\} = (1 - 1/T)$ and (5.43) follows. □

Comparing (5.22) with (5.43), we deduce that hard collisions in OCRA happen with exactly the same frequency as in non-cooperative SSRA. This is a design goal made possible by the increase in the PN sequence period $\mathcal{P}$ as per [S0]. Certainly, this period cannot be made arbitrarily large since it must satisfy $\mathcal{P} \leq 2^S$, [24], effectively limiting the maximum achievable diversity order of OCRA to

$$\kappa = \frac{2^S - 1}{T}. \tag{5.46}$$

Notice though that since in general $2^S/T \gg 1$, the constraint in (5.46) is not severe in practice.

[2]This requires noting that the sum of two normal random variables is also normally distributed so that the fading of the "combined" diversity path is also Rayleigh; see also (5.52).

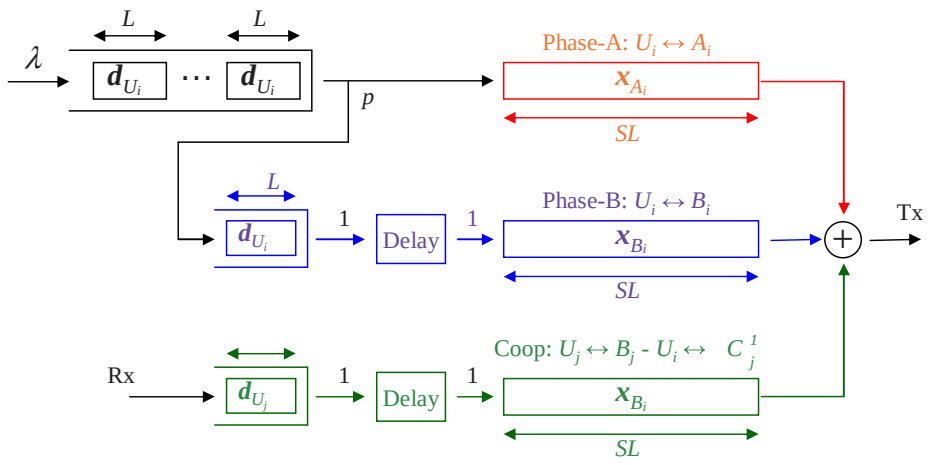

Figure 5.6: Each terminal has three independent transmission chains that are combined using baseband digital signal processing.

To wrap up this section, let us look at OCRA from the perspective of a terminal; see also Fig. 5.6. Each terminal maintains three separate transmission chains: the first one for the transmission of phase-A packets, a second one for the transmission of phase-B packets and a third one for the transmission of cooperative packets. The phase-A chain is used with probability p ([S1]) and is fed with packets from the terminal's queue. If the user was in phase-A during the previous slot then it enters phase-B in the current one, activating the second transmission chain to transmit the packet stored in the phase-A buffer. The third chain is used when cooperating with other users and is activated whenever a packet is successfully decoded during the idle state.

The terminal can use more than one chain simultaneously, if it decides to enter phase-A in two consecutive slots, or, if it decodes another terminal's packet in the slot immediately before entering phase-A. Interestingly, not all the chains can be used simultaneously. As we can see from Fig. 5.5 mixed states include active-A plus cooperator and active-A plus active-B. Mixed states including active-B and cooperator never happen since this would require decoding a packet (to become cooperator) and being active-A (to become active B) in the previous slot. This is impossible for half-duplex terminals and consequently the active-B and cooperation chains are never used simultaneously.

Remark 17 This multi-transmission ability ensures that at any given time the random variables N_A and N_B are not only independent of each other but also that their distribution is not affected by the cooperation among users. Assuming a saturated system, we have that N_A and N_B follow binomial distributions with parameters J and p; i.e.,

$$\Pr\{N_A = n\} = \Pr\{N_B = n\} = \binom{J}{n} p^n (1-p)^{J-n}. \tag{5.47}$$

Beyond a saturated system, this expression is also valid for the dominant system (see Section 5.4). Finally, note that if U_j enters phase-A while being active-B or cooperator, it will fail in recruiting cooperators with high probability due to the self interference from high-power phase-B packets to low-power phase-A packets. This rather undesirable situation should be avoided in practice, but is allowed here to ensure independence between N_A and N_B.

5.3.1 Packet transmission and reception

The first problem we consider is signal transmission and reception in OCRA to abide by [S0]-[S7]. There are two signal reception instances in OCRA that we have to study. One is the detection of phase-A packets by nearby idle users and the other one is the detection of the cooperative transmission of phase-B packets. If we call $\mathbf{d}_{A_j} = \{d_{A_j}(l)\}_{l=0}^{L-1}$ the unit-power information packet of the active-A user A_j, then the corresponding transmitted packet $\mathbf{x}_{A_j}$ is constructed according to [S2] with entries

$$x_{A_j}(Sl+s) = \sqrt{P(A_j)}\; d_{A_j}(l) c(Sl+s), \qquad l \in [0, L-1],\ s \in [0, S-1], \tag{5.48}$$

where we used $\tau_{A_j} = 0$ and $P(A_j)$ is given by (5.40). Likewise, if $\mathbf{d}_{B_j} = \{d_{B_j}(l)\}_{l=0}^{L-1}$ is the packet of the active-B user B_j, the packet transmitted by a given cooperator C_j^k is constructed according to [S5] and given by

$$x_{C_j^k}(Sl+s) = \sqrt{P(C_j^k)} d_{B_j}(l) c(Sl+s-\tau_{C_j^k}),\ k \in [1, K_j] \tag{5.49}$$

with $\tau_{C_j^k}$ and $P(C_j^k)$ as in (5.41).

We first analyze the reception of a packet from a reference active-B user B_{j_0}. For that matter, let the received block at the AP be $\mathbf{z} = \{z(t)\}_{t=0}^{T-1}$ whose components are given by

$$z(Sl+s) = \sum_{j=1}^{N_B} d_{B_j}(l) \sum_{k=0}^{K_j} \sqrt{P(C_j^k)}\; h(C_j^k) c(Sl+s-\tau_{C_j^k}) + \sum_{j=1}^{N_A} \sqrt{P(A_j)}\; h(A_j) d_{A_j}(l) c(Sl+s) + n(Sl+s) \tag{5.50}$$

that is, the superposition of the cooperative N_B active-B transmissions, the N_A low power active-A transmissions and the receiver noise.

Let us focus on the detection of any one of the diversity paths of B_{j_0}'s communication say the one with PN-shift $\tau_{B_{j_0},\kappa_0} := \tau_{B_{j_0}} + \kappa_0 T$. Since according to [S5] this shift is chosen by a random number of cooperators, we define the number of B_{j_0}'s cooperators that chose this shift as

$$N(B_{j_0}, \kappa_0) := \#\{C_{j_0}^k \in \mathcal{C}_{j_0} \text{ s.t. } \tau_k = \kappa_0\} := \#(C_{j_0}^{\kappa_0}) \tag{5.51}$$

where the cardinality operator # represents the number of elements in a set. Since the packets $\mathbf{x}_{C_{j_0}^k}$ of all cooperators in the set $C_{j_0}^{\kappa_0}$ share the PN shift $\tau_{B_{j_0}} - \kappa_0 T$, they are indistinguishable at the AP. Thus, all cooperators in $C_{j_0}^{\kappa_0}$ in (5.51) appear as a single path to the AP with composite Rayleigh fading coefficient

$$h(C_{j_0}^{\kappa_0}) := \sum_{k:\tau_k=\kappa_0} P(C_{j_0}^k) h(C_{j_0}^k). \tag{5.52}$$

Note that being a sum of complex Gaussian random variables, $h(C_{j_0}^{\kappa_0})$ is also complex Gaussian and the composite fading is also Rayleigh.

To recover the path $C_{j_0}^{\kappa_0}$, the AP compensates for the random phase by multiplying with the normalized composite channel conjugate $h_n^*(C_{j_0}^{\kappa_0}) := h^*(C_{j_0}^{\kappa_0})/|h(C_{j_0}^{\kappa_0})|$ and despreads with the proper PN shift. This yields the decision vector $\mathbf{r}_{C_{j_0}^{\kappa_0}} = \{r_{C_{j_0}^{\kappa_0}}(l)\}_{l=0}^{L-1}$ with entries

$$r_{C_{j_0}^{\kappa_0}}(l) = h_n^*(C_{j_0}^{\kappa_0}) \frac{1}{S} \sum_{s=0}^{S-1} z(Sl+s) c(Sl+s-\tau_{B_{j_0}} - \kappa_0 T). \tag{5.53}$$

If a hard collision does not occur, then $\tau_{B_{j_0}} \neq \tau_{B_j}\ \forall\ j \neq j_0$ and straightforward manipula-

tions (see Appendix A.1) yield the per-path SINR as:

$$\begin{aligned}\mathrm{SINR}(B_{j_0},\kappa_0) &:= \frac{\mathrm{E}^2[r_{C_{j_0}^{\kappa_0}}(l)]}{\mathrm{var}[r_{C_{j_0}^{\kappa_0}}(l)]} \\ &= S\frac{N(B_{j_0},\kappa_0)/(K_{j_0}+1)}{(N_B-1)+\left(1-\frac{N(B_{j_0},\kappa_0)}{K_{j_0}+1}\right)+\rho N_A+N_0/P_0}.\end{aligned} \tag{5.54}$$

Coherent combining of these κ_j paths leads to diversity order κ_j, with the PEP determined by the $\mathrm{SINR}(B_{j_0},\kappa_0)$ given by (5.54) for all the shifts $\kappa_0 \in [0,\kappa]$. Note that the denominator of $\mathrm{SINR}(B_{j_0},\kappa_0)$ in (5.54) contains a term $(N_B-1)P_0$ accounting for the interference from other active-B users, a term $[1-N(B_{j_0},\kappa_0)/(K_{j_0}+1)]P_0$ accounting for the self-interference of other paths of the same communication $B_{j_0} \to AP$ and a term $\rho N_A P_0$ for the active-A users's interference.

Remark 18 The analysis in this section should clarify the difference between cooperation order and diversity order as defined in [S5]. Note that κ_j is indeed the diversity order of the $B_j \to AP$ link, since the number of uncorrelated Rayleigh channels is precisely κ_j. In that regard, OCRA's diversity depends not only on the number of cooperation order K_j – as usual in most cooperative protocols – but also on the (random) selection of PN shifts by the user in $\mathcal{C}_j$.

The other reception instance is that of idle users decoding active-A transmissions. Consider the received vector at the idle user I_i denoted by $\mathbf{z}_{I_i} = \{z_{I_i}(t)\}_{t=0}^{T-1}$ with entries

$$\begin{aligned}z_{I_i}(Sl+s) =& \sum_{j=1}^{N_B}\sum_{k=0}^{K_j}\sqrt{P(C_j^k)}h(C_j^k,I_i)d_{B_j}(l)c[Sl+s-\tau_{C_j^k}] \\ &+\sum_{j=1}^{N_A}\sqrt{P(A_j)}h(A_j,I_i)d_{A_j}(l)c(Sl+s)+n(Sl+s).\end{aligned} \tag{5.55}$$

In this case, we focus on decoding the reference active-A user $U_0 = A_{N_A}$. To construct the pertinent decision variable, we have to compensate for fading by multiplying with $h_n^*(U_0,I_i) := h^*(U_0,I_i)/|h(U_0,I_i)|$ and despreading with $c(t)$. Letting $\mathbf{r}_{I_i} = \{r_{I_i}(l)\}_{l=0}^{L-1}$ be the decision vector, we have

$$r_{I_i}(l) \;=\; h_n^*(U_0,I_i)\;\frac{1}{S}\sum_{s=0}^{S-1} z_{I_i}(Sl+s)c(Sl+s). \tag{5.56}$$

As we did for the AP, we can obtain the mean and variance of $r_{U_0}(l)$ (see Appendix A.2), and from there SINR_0^i, the SINR at idle user I_i for the signal of U_0. Its inverse is given by

$$(\mathrm{SINR}_0^i)^{-1} := \frac{\mathrm{var}[r_{I_i}(l)]}{\mathrm{E}^2[r_{I_i}(l)]} = S^{-1}\sum_{j=1}^{N_B}\sum_{k=0}^{K_j}\frac{P(C_j^k \to I_i)}{P(U_0 \to I_i)} + \sum_{j=1}^{N_A-1}\frac{P(A_j \to I_i)}{P(U_0 \to I_i)} + S^{-1}\frac{N_0}{P(U_0 \to I_i)} \tag{5.57}$$

where the powers $P(U_j \to I_i)$ for the different users are obtained from the path loss model in (5.3).We remark that the interference from other active-A users is not reduced by the spreading gain, but (hopefully) by spatial separation.

The SINR in (5.57) determines the probability of I_i becoming a cooperator of U_0, and as such, it is an important metric of OCRA that we will study in Section 5.5. But before that, we will introduce our main result pertaining to OCRA's throughput.

5.4 OCRA's throughput

Mimicking the steps we followed for the non-cooperative SSRA protocol in Section 5.2, we can try to evaluate the aggregate throughput of OCRA. The hard collision probability coincides with the non-cooperative SSRA protocol and is given by Proposition 5. The soft collision probability, on the other hand, depends on both the number of active-A and active-B users and is given by [c.f. (5.23)]

$$P_{\mathrm{SC}}(N_A, N_B) = P_e(N_A, N_B)[1 - P_{\mathrm{HC}}(N_B)], \tag{5.58}$$

with $P_e(N_A, N_B)$ a function that maps the number of active-A and active-B users to the average PEP.

Using (5.58), we can compute the packet success probability conditioned on the number of interferers, namely $P_s(N_A, N_B) := 1 - P_{\mathrm{HC}}(N_B) - P_{\mathrm{SC}}(N_A, N_B))$. Using the latter and (5.43), (5.58) we find

$$P_s(N_A, N_B) = \left(1 - \frac{1}{T}\right)^{N_B}[1 - P_e(N_A, N_B)]. \tag{5.59}$$

Averaging (5.59) over the joint distribution of (N_A, N_B) and considering the average departure rate definition in (5.25), we find

$$\mu^{\mathrm{OCRA}} = p \sum_{n_B=0}^{J-1} \Pr\{N_B = n_B\} \left(1 - \frac{1}{T}\right)^{n_B} \times \sum_{n_A=0}^{J} \Pr\{N_A = n_A\}[1 - P_e(n_A, n_B + 1)] \tag{5.60}$$

were we used the independence of N_A and N_B discussed in Remark 17. For a saturated system, the probabilities $\Pr\{N_B = n_B\}$ and $\Pr\{N_A = n_A\}$ are binomially distributed as in (5.47). This motivates introduction of the dominant system obtained after replacing [S1] with:

[S1'] At the beginning of each slot, U_j enters phase-A with probability p and moves the first packet in its queue, $\mathbf{d}_{U_j} := \{d_{U_j}(l)\}_{l=0}^{L-1}$, to the phase-A buffer. If U_j's queue is empty, it moves a dummy packet.

This modification renders the departure process stationary and we can claim, as we did in the proof of Proposition 4, that $\eta^{\mathrm{OCRA}} = J\mu^{\mathrm{OCRA}}$, with μ^{OCRA} given as in (5.60).

The difficulty in evaluating OCRA's throughput is cocooned in the function $P_e(N_A, N_B)$. This function depends on the diversity order κ_j, which depends on the number of cooperators K_j recruited during phase-A; while in theory we could compute K_j's distribution and from there $P_e(N_A, N_B)$, this turns out to be analytically intractable and motivates the asymptotic approach of the next section.

5.4.1 OCRA's asymptotic throughput

Since OCRA's throughput $\eta^{\mathrm{OCRA}}(J, N_0/P_0, S, \kappa, p, \rho)$ depends also on (κ, ρ), it is convenient to differentiate the MST (as defined in (5.32)) depending on whether we optimize over ρ or not. If we consider ρ fixed, we define the ρ-conditional MST as:

$$\eta_{\max}^{OCRA}(J, N_0/P_0, S, \kappa|\rho) = \max_{p} \{\eta^{OCRA}(J, N_0/P_0, S, p, \kappa, \rho)\} \tag{5.61}$$

with the maximum achieved at $p_{\max}(\rho) = \arg\max_p(\eta)$. If we jointly optimize over (p, ρ), we define the MST as:

$$\eta_{\max}^{OCRA}(J, N_0/P_0, S, \kappa) = \max_{p,\rho} \left\{\eta^{OCRA}(J, N_0/P_0, S, p, \kappa, \rho)\right\} \tag{5.62}$$

with the maximum achieved at $(p_{\max}, \rho_{\max}) = \arg\max_{(p,\rho)}(\eta)$. We adopt this second definition as the one equivalent to the non-cooperative SSRA MST defined in (5.32).

Having made this distinction, we can introduce the main results of this chapter in the following two theorems.

Theorem 8 *Consider the OCRA dominant system defined by rules* [S0], [S1'] *and* [S2]-[S7] *operating over a fading channel; and functions* $\rho = \rho(J)$ *and* $K = K(J)$ *such that* $\lim_{J\to\infty} \rho = 0$ *and* $\lim_{J\to\infty} K = \infty$. *Let* $\mathcal{C}_j := \{C_j^k\}_{k=1}^{K_j}$ *be the set of cooperators of the active-B user* B_j *for* $j \in [1, N_B]$. *If*

[**h1**] $\lim_{J\to\infty}(\rho^{2/\alpha} J/K) = \infty$, *with* α *being the pathloss exponent in (5.3); and*

[**h2**] *the transmission probability* $p = p_{\max}(\rho)$ *is chosen to achieve the MST given* ρ;

then

$$\lim_{J\to\infty} \Pr\{K_j \geq K/2, \ \ \forall j\} = 1. \tag{5.63}$$

Proof: See Section 5.5.2.

Theorem 8 establishes that every active-B user is receiving cooperation by at least $K/2$ users; moreover, as long as the convergence rates of $\rho(J)$ and $K(J)$ satisfy [h1] the cooperation order K_j becomes arbitrarily large while the active-A transmitted power becomes arbitrarily small. Consequently, the seemingly conflicting requirements of recruiting an infinite number of cooperators with a vanishingly small power *are* compatible as $J \to \infty$ implying that very large diversity orders are achievable by OCRA. A by-product of this comment leads to the following result.

Theorem 9 *For any* $\kappa \leq (2^S - 1)/T$, *the asymptotic MST of OCRA operating over a Rayleigh fading channel* η_∞^{OCRA} *and the asymptotic throughput of non-cooperative random*

access over a κ-order, diversity channel η_∞^κ are equal; i.e.,

$$\lim_{J\to\infty} \eta_{\max}^{OCRA}(J, N_0/P_0, S, \kappa) := \eta_\infty^{OCRA}(N_0/P_0, S, \kappa)$$
$$= \eta_\infty^\kappa(N_0/P_0, S). \tag{5.64}$$

Proof: For each value of J, choose (K, ρ, p) according to the conditions of Theorem 8. With its hypotheses satisfied, Theorem 8 states that for any active-B user we can map an arbitrarily large ($K_j > K/2$) number of cooperators to a finite number of PN shifts κ. Accordingly, the number of elements in the set $C_j^{\kappa_0}$ in (5.51) satisfies

$$\lim_{J\to\infty} \frac{N(B_j, \kappa_0)}{K_j + 1} = 1/\kappa, \quad \forall\ \kappa_0 \in [1, \kappa], \tag{5.65}$$

due to the law of large numbers. Using (5.65) and $\lim_{J\to\infty} \rho = 0$, the per path SINR in (5.54) reduces to

$$\lim_{J\to\infty} \mathrm{SINR}(B_j, \kappa_0) = S\frac{1/\kappa}{N_B - 1 + (1 - 1/\kappa) + N_0/P_0}$$
$$:= \bar{\gamma}(N_B, \kappa). \tag{5.66}$$

Eq. (5.66) is, in part, a manifestation of the fact that as $J \to \infty$, the active-A users transmit with negligible power. But note that (5.66) is identical to the per-path SINR in a κ-order diversity channel [c.f. (5.36)], and because it is valid for every active-B user B_j and every shift κ_0 we infer that

$$\lim_{J\to\infty} P_e(N_A, N_B) = P_e^\kappa(\kappa\bar{\gamma}(N_B, \kappa)), \tag{5.67}$$

with $P_e(N_A, N_B)$ the function determining $P_{\mathrm{SC}}(N_A, N_B)$ in (5.58) and $P_e^\kappa(\kappa\bar{\gamma}(N_B, \kappa)) = P_e^\kappa(\bar{\gamma}_{N_B})$ the corresponding member of the family of functions introduced in Definition 7.

Even though computing $P_e(N_A, N_B)$ is intractable, we can find its limit as $J \to \infty$; moreover, in the limit $P_e(N_A, N_B)$ is a function of N_B only, and we can compute the limit of the average departure rate in (5.60) as

$$\lim_{J\to\infty} \mu_{\max}^{\mathrm{OCRA}} = p_{\max} \sum_{n_B=0}^{J-1} \binom{J-1}{n_B} p_{\max}^{n_B} (1 - p_{\max})^{J-1-n_B}$$
$$\times \left(1 - \frac{1}{T}\right)^{n_B} [1 - P_e^\kappa(\kappa\bar{\gamma}(N_B, \kappa))] \tag{5.68}$$

This is identical to the expression (5.28) of Proposition 4 when the channel is a κ-order diversity channel establishing that $\lim_{J\to\infty} J\mu_{\max}^{\mathrm{OCRA}} = \eta_{\infty}^{\kappa}(N_0/P_0, S)$. To complete the proof, we invoke the same argument used in Proposition 4 about the dominant system to claim that

$$\begin{aligned}\eta_{\infty}^{OCRA}(N_0/P_0, S, \kappa) &:= \lim_{J\to\infty} \eta_{\max}^{OCRA}(J, N_0/P_0, S, \kappa) \\ &= \lim_{J\to\infty} J\mu_{\max}^{\mathrm{OCRA}} \\ &= \eta_{\infty}^{\kappa}(N_0/P_0, S). \end{aligned} \tag{5.69}$$

The first equality follows form the definition of asymptotic throughput in (5.64), the second from the dominant system argument, and the last one by comparing (5.68) with (5.28). □

Theorem 9 is the main result of this chapter effectively stating that very high diversity orders are achievable by OCRA. Notice that the only constraint $\kappa \leq (2^S - 1)/T$, is not very restrictive in practice since we are interested in achieving diversity orders of no more than a few units and $2^S/T \gg 1$. Thus, it is fair to recall Remark 14 and assert that

$$\eta_{\infty}^{OCRA}(N_0/P_0, S, \kappa) = \eta_{\infty}^{\kappa}(N_0/P_0, S) \approx \eta_{\infty}^{G}(N_0/P_0, S), \tag{5.70}$$

with κ sufficiently large.

Surprisingly, user cooperation can improve the network throughput to the point of achieving wireline-like throughput in a wireless RA environment. This is a subtle but significant difference relative to point-to-point user cooperation in fixed access networks, where the diversity advantage typically comes at the price of bandwidth expansion [51, 96].

5.5 On the asymptotic behavior of OCRA

In this section, we will show that Theorem 8 is a consequence of the spatial distribution of users. We will first consider a particular snapshot of an OCRA system with arbitrarily large N_I but fixed N_A and N_B, and study the distance ratios that determine the SINR (Lemma 5). From there, we prove that if [h1] in Theorem 8 is true, then every $I_0^{(k)}$ with $k \leq K$ correctly decodes U_0's phase-A packet almost surely (Theorem 10). We will then

establish that with high probability, the numbers of users N_A, N_B, and N_I in OCRA behave like the numbers of this particular snapshot (Lemma 6) from where Theorem 8 will follow.

5.5.1 A network snapshot

We consider in this subsection a fixed access network corresponding to a snapshot of the OCRA dominant system operating under [S0], [S1'] and [S2]-[S7]. In this fixed access network, N_A and N_B are fixed but the number of idle users $N_I \to \infty$. The problem we are concerned with is that of the reference active-A user U_0 trying to communicate with the idle users in $\mathcal{I}$. For each member of $\mathcal{I}$, the detection probability is determined by the SINR. If we let $\mathrm{SINR}_0^{(k)}$ be such a metric at the k^{th} closest to U_0 idle user, we have

$$(\mathrm{SINR}_0^{(k)})^{-1} := S^{-1} \sum_{j=1}^{N_B} \sum_{k=0}^{K_j} \frac{P(C_j^k \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})} + \sum_{j=1}^{N_A-1} \frac{P(A_j \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})} + \frac{N_0}{P(U_0 \to I_0^{(k)})} \tag{5.71}$$

which is obtained by setting $I_i = I_0^{(k)}$ in (5.57). The first sum in (5.71) corresponds to the N_B active-B users, the second sum to the N_A active-A users, and the third term accounts for the receiver noise. The upper limit $N_A - 1$ of the second sum follows from the convention $U_0 = A_{N_A}$.

To relate power terms in (5.71) with corresponding distances, let us consider first the (interfering) power received at I_0^k from B_j's communication which involves the set of $K_j + 1$ cooperators $\mathcal{C}_j = \{C_j^k\}_{k=0}^{K_j}$:

$$P(B_j \to I_0^{(k)}) := \sum_{i=0}^{K_j} P(C_j^i \to I_0^{(k)}) = \frac{1}{K_j+1} \sum_{i=0}^{K_j} \frac{P_0 \|C_j^i\|^\alpha}{\|C_j^i - I_0^{(k)}\|^\alpha} \tag{5.72}$$

where the second equality comes from the path loss model in (5.3) and the average power control enacted by [S5]. Less severe but not negligible interference is received from active-A users; for a specific A_j, we have

$$P(A_j \to I_0^{(k)}) = \frac{\rho P_0 \|A_j\|^\alpha}{\|A_j - I_0^{(k)}\|^\alpha}. \tag{5.73}$$

Remembering that U_0's phase-A power is $\rho P_0 \|U_0\|^\alpha/\xi$ (so that it is received at the AP with power ρP_0), the signal power received at $I_0^{(k)}$ is $P(U_0 \to I_0^{(k)}) = \rho P_0 \|U_0\|^\alpha / \|U_0 - I_{(0)}^{(k)}\|^\alpha$, and we obtain [c.f. (5.71), (5.72), and (5.73)]

$$(\mathrm{SINR}_0^{(k)})^{-1} = \frac{1}{S\rho} \sum_{j=1}^{N_B} \sum_{i=0}^{K_j} \frac{1}{K_j+1} \frac{\|C_j^i\|^\alpha \|U_0 - I_0^{(k)}\|^\alpha}{\|U_0\|^\alpha \|C_j^i - I_0^{(k)}\|^\alpha} + \sum_{j=1}^{N_A-1} \frac{\|A_j\|^\alpha \|U_0 - I_0^{(k)}\|^\alpha}{\|U_0\|^\alpha \|A_j - I_0^{(k)}\|^\alpha} + \frac{N_0}{\rho P_0} \frac{\|U_0 - I_0^{(k)}\|^\alpha}{\|U_0\|}. \tag{5.74}$$

The SINR expression in (5.74) determines the probability that a packet transmitted by U_0 with reduced power ($\rho \ll 1$) is received correctly at the k^{th} closest to U_0 idle user. We would prefer $\rho \to 0$ so that the interference added to the AP in (5.54) is negligible, and we want $k \to \infty$ so that the cooperation order grows large. As commented before, it will turn out that these seemingly conflicting requirements *are* compatible for N_I sufficiently large.

To establish this we need to establish two lemmas; the first one concerns the cumulative distribution function (CDF) of the distance between any two users.

Lemma 4 *If users are uniformly distributed in a disc of radius R, U_j denotes an arbitrary user (idle, active-A or active-B), and $F(r) := \Pr\{\|U_j - U_0\| < r | U_0\}$, then $F(r) = 0$ for $r < 0$, and*

$$\min\left\{\frac{r^2}{4R^2}, 1\right\} \le F(r) \le \min\left\{\frac{r^2}{R^2}, 1\right\}, \text{ for } r > 0. \tag{5.75}$$

Proof: See Appendix A.

Since users are uniformly distributed within a circle, their distance $\|U_j\|$ to the AP follows a quadratic CDF as asserted by (5.1). Lemma 4 establishes that their distance to any point, in this case to the reference user U_0, has a CDF that is lower and upper bounded by a parabola.

This result is useful in establishing that some pertinent distance ratios are becoming arbitrarily large, as we quantify in the next lemma.

Lemma 5 *With N_I denoting the number of idle users, consider a function $\rho = \rho(N_I)$ that determines the phase-A fraction of power and a function $K = K(N_I)$ such that* $\lim_{N_I \to \infty} \rho =$

0, $\lim_{N_I\to\infty} K = \infty$ *and* $\lim_{N_I\to\infty}(\rho^{2/\alpha} N_I/K) = \infty$. *Then, for arbitrary* $\mathcal{K} > 0$, *the events*

$$e_1(N_I, \mathcal{K}) := \{\|U_0\| > (\mathcal{K}/\rho^{1/\alpha})\|I_0^{(K)} - U_0\|\} \tag{5.76}$$

$$e_2(N_I, \mathcal{K}) := \{\|B_0^{(1)} - U_0\| > (\mathcal{K}/\rho^{1/\alpha})\|I_0^{(K)} - U_0\|\} \tag{5.77}$$

$$e_3(N_I, \mathcal{K}) := \{\|A_0^{(1)} - U_0\| > \mathcal{K}\|I_0^{(K)} - U_0\|\} \tag{5.78}$$

have probability 1 as the number of idle users $N_I \to \infty$; *i.e.,*

$$\lim_{N_I\to\infty} \Pr\{e_l(N_I, \mathcal{K})\} = 1, \qquad l = 1, 2, 3. \tag{5.79}$$

Proof: See Appendix B.

If we let $\xrightarrow{p}$ denote convergence in probability, Lemma 5 implies that the distance ratios satisfy

$$\frac{\rho^{1/\alpha}\|U_0\|}{\|I_0^{(K)} - U_0\|}, \quad \frac{\rho^{1/\alpha}\|B_0^{(1)} - U_0\|}{\|I_0^{(K)} - U_0\|}, \quad \frac{\|A_0^{(1)} - U_0\|}{\|I_0^{(K)} - U_0\|} \xrightarrow{p} \infty, \tag{5.80}$$

for every ρ and K satisfying the conditions of Lemma 5.

Intuitively, U_0's phase-A transmission will not be correctly decoded by $I_0^{(K)}$ when compared to the distance $\|I_0^{(K)} - U_0\|$, either because $I_0^{(K)}$ is close to an active-B user, or close to another active-A user, or, because U_0 is close to the AP. In the first two cases, the interference will be too high, and in the third case the signal will be too weak (being close to the AP, the power $P(U_0)$ is small because of [S2]). The importance of Lemma 5 is in establishing that all these events happen with vanishing probability and points out to the almost certainty of $I_0^{(K)}$ decoding U_0's phase-A transmission successfully. This is formally asserted in the following theorem.

Theorem 10 *Consider a set of* N_A *active-A users,* $\mathcal{A} := \{A_j\}_{j=1}^{N_A}$; *a set of* N_I *idle users,* $\mathcal{I} := \{I_j\}_{j=1}^{N_I}$; *and a set of* N_B *active-B users,* $\mathcal{B} := \{B_j\}_{j=1}^{N_B}$, *each receiving cooperation from a set of* K_j *idle users,* $\mathcal{C}_j := \{C_j^k\}_{k=1}^{K_j}$. *Let* $U_0 = A_{N_A}$ *be a reference user,* $I_0^{(k)}$ *be the* k^{th} *closest to* U_0 *idle user and* $\mathcal{C}_0 := \{C_0^k\}_{k=0}^{K_0}$ *be the set of idle users that decode* U_0*'s phase-A packet correctly (called* U_0*'s cooperators). If*

[h1] *the functions* $\rho = \rho(N_I)$ *and* $K = K(N_I)$ *satisfy* $\lim_{N_I\to\infty} \rho = 0$ *and* $\lim_{N_I\to\infty} K = \infty$;

[h2] *convergence rates are such that* $\lim_{N_I\to\infty}(\rho^{2/\alpha}N_I/K)=\infty$*; and*

[h3] *the transmitted powers are* $P(C_j^k)=P_0\|C_j^k\|^\alpha/[\xi(K_j+1)]$, $P(A_j)=(\rho P_0/\xi)\|A_j\|^\alpha$ *and* $P(U_0)=(\rho P_0/\xi)\|U_0\|^\alpha$*;*

then

[a] *as* $N_I\to\infty$*, the ratio of distances between* B_j *and its farthest cooperator* $C_j^{(K_j)}$ *and the distance between* B_j *and the AP converges to* 0 *in probability; i.e.,*

$$\lim_{N_I\to\infty}\Pr\left\{\frac{\|B_j-C_j^{(K_j)}\|}{\|B_j\|}<\epsilon\right\}=1,\ \forall\ \epsilon>0 \tag{5.81}$$

[b] *for every* $k\le K$*, the event that* $I_0^{(k)}$ *becomes a cooperator is asymptotically almost sure; i.e.,*

$$\lim_{N_I\to\infty}\Pr\{I_0^{(k)}\in\mathcal{C}_0\}=1. \tag{5.82}$$

Proof: See Appendix D.

Theorem 10-[a] states that as we reduce the phase-A fraction of power, we do not recruit faraway idle users. In that sense, cooperators become clustered around the active-B user they are cooperating with nicely matching the intuition of cooperation with nearby users.

More important, Theorem 10-[b] establishes that the probability of each $I_0^{(k)}$, $k\le K$, becoming a cooperator when phase-A transmission is reduced by a factor ρ converges to 1, as the number of idle users N_I grows large. Moreover, as long as $\lim_{N_I\to\infty}(\rho^{2/\alpha}N_I/K)=\infty$, the phase-A fraction of power ρ can be made arbitrarily small and the number K of cooperators recruited arbitrarily large. The mathematical formalism here should not obscure the fact that this suggests the possibility of having an arbitrarily large number of terminals correctly decoding U_0's active-A transmission with probability 1; and correspondingly enable arbitrarily large diversity order during phase-B when U_0 transmits with practically negligible power during phase-A.

Applying Theorem 10 to OCRA requires taking care of the randomness in the number of active-A and active-B users in a given slot, a problem that leads us to the next section.

5.5.2 Asymptotic Throughput

Theorem 10 establishes the potentially high cooperation order of the described fixed network access. The following lemma establishes that with high probability, an OCRA network is well described by the fixed network for which Theorem 10 has been proved.

Lemma 6 *Let $p_{\max}$ be the probability that achieves MST of the OCRA dominant system defined by rules* [S0], [S1'] *and* [S2]-[S7]*; assume that $0 < \eta_\infty := \lim_{J\to\infty} \eta_{\max} < \infty$ exists; and let $\bar{N} := \mathrm{E}(N_A) = \mathrm{E}(N_B) = p_{\max} J$ denote the average number of active-A (active-B) users. It then holds that*

[a] the average number of users converges

$$\lim_{J\to\infty} \bar{N} = \bar{N}_\infty \tag{5.83}$$

to a finite constant $\bar{N}_\infty \in (0, \infty)$; and,

[b] the random variables N_A and N_B are asymptotically Poisson distributed:

$$\Pr\{N_B = n\} = \Pr\{N_A = n\} = \frac{\bar{N}^n}{n!} e^{-\bar{N}}. \tag{5.84}$$

Proof: If $\bar{N} \to \infty$, then the probability that *all* active-B users experience a hard collision goes to 1:

$$\lim_{J\to\infty} \bigcap_{j_0} \left\{ \bigcup_{j\neq j_0} \left\{ \tau_{U_{j_0}} = \tau_{U_j} \right\} \right\} = 1 \,, \tag{5.85}$$

since we have a finite number of PN shifts T and an infinite number of instantaneously active users; thus, $\lim_{J\to\infty} \bar{N} \neq \infty$. The fact that $\bar{N}$ does not oscillate follows since $\bar{N}(J)$ is a non-decreasing function of J from where (5.83) follows. To prove claim [b], simply note that the conditions of Poisson's theorem are satisfied. □

The importance of Lemma 6 is in establishing that as $J \to \infty$, the average number of active-A (active-B) users remains bounded; i.e., $\bar{N} \to \bar{N}_\infty < \infty$. This enables application of Theorem 10 to establish the asymptotically infinite order diversity of the OCRA network as claimed by Theorem 8 that we are now ready to prove.

Proof of Theorem 8: Eq. (5.63) can be written in terms of the complementary event

$$\Pr\{\cup_j(K_j < K/2)\} = 1 - \Pr\{K_j \geq K/2 \ \ \forall j\}, \tag{5.86}$$

which we will prove convergent to zero. To this end, let us start by defining a network snapshot as the set $\mathcal{S} := \{\mathcal{U}, \mathcal{A}, \mathcal{B}, \cup_{j|U_j\in\mathcal{B}}\mathcal{C}_j\}$ composed of the realizations of user's positions and classes; and the index $k^* = \arg\max_{k\in[1,K]} \Pr\{I_0^{(k)} \notin \mathcal{C}_0|\mathcal{S}\}$ corresponding to the idle user least likely to decode U_0 among the K closest ones when the snapshot $\mathcal{S}$ is given.

We separate the failure in soliciting at least $K/2$ cooperators – the event $\{\cup_j(K_j < K/2)\}$ in (5.86) – in two cases: i) the realization $\mathcal{S}$ is not favorable and we fail with high probability, e.g., when N_A, N_B are very large; and ii) $\mathcal{S}$ is favorable and we succeed with high probability. For that matter, define the set of network realizations $\mathcal{S}_{\beta,N\max} := \{\mathcal{S}|\Pr\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\} \leq \beta; N_A, N_B \leq N_{\max}\}$ for which the number of active-A and active-B users is less than $N_{\max}$, and the decoding failure probability is less than β, to write

$$\begin{aligned}\Pr\{\cup_j(K_j<K/2)\} =& \Pr\{\cup_j(K_j<K/2)|\mathcal{S}_{\beta,N_{\max}}\}\Pr\{\mathcal{S}_{\beta,N_{\max}}\}\\ &+\Pr\left\{\cup_j(K_j<K/2)|\overline{\mathcal{S}_{\beta,N_{\max}}}\right\}\Pr\left\{\overline{\mathcal{S}_{\beta,N_{\max}}}\right\}.\end{aligned} \tag{5.87}$$

Further recalling that probabilities are smaller than 1 we obtain

$$\begin{aligned}\Pr\{\cup_j(K_j<K/2)\} \leq& \Pr\{\cup_j(K_j<K/2)|\mathcal{S}_{\beta,N_{\max}}\}\\ &+ \Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}\}.\end{aligned} \tag{5.88}$$

Applying the union bound to the event $\{\cup_j(K_j < K/2)|\mathcal{S}_{\beta,N_{\max}}\}$, we obtain

$$\begin{aligned}\Pr\{\cup_j(K_j < K/2)\} \leq& N_{\max}\Pr\{K_j < K/2|\mathcal{S}_{\beta,N_{\max}}\}\\ &+ \Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}\}\end{aligned} \tag{5.89}$$

since the number of active-B users is $N_B \leq N_{\max}$.

We start by bounding the first term in (5.89). To this end, we note that in order for $K_j < K/2$ we must have at least $K/2$ decoding failures among the K closest idle users during phase-A. Furthermore, the decoding probabilities at idle users are independent

when conditioned on the network snapshot $\mathcal{S}$; i.e., $\Pr\{I_0^{(k_1)}, I_0^{(k_2)} \in \mathcal{C}_0|\mathcal{S}\} = \Pr\{I_0^{(k_1)} \in \mathcal{C}_0|\mathcal{S}\}\Pr\{I_0^{(k_2)} \in \mathcal{C}_0|\mathcal{S}\}$, and we can thus write

$$\Pr\{K_j < K/2|\mathcal{S}\} < \sum_{k=K/2}^{K} \binom{K}{k} \left[\Pr\left\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\right\}\right]^k \times \left[\Pr\left\{I_0^{(k^*)} \in \mathcal{C}_0|\mathcal{S}\right\}\right]^{K-k} \tag{5.90}$$

where we used the fact that by definition $\Pr\{I_0^{(k)} \notin \mathcal{C}_0|\mathcal{S}\} \leq \Pr\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\}$ for all $k \in [1, K]$. The largest summand in (5.90) corresponds to $k = K/2$, which together with $\Pr\{I_0^{(K)} \in \mathcal{C}_0\}^{K-k} < 1$, yields

$$\begin{aligned}\Pr\{K_j < K/2|\mathcal{S}\} &< K/2\binom{K}{K/2}\left[\Pr\left\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\right\}\right]^{K/2} \\ &\leq (K/2)2^K\left[\Pr\left\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\right\}\right]^{K/2}\end{aligned} \tag{5.91}$$

where we also used Stirlings' factorial approximation to obtain the last expression.

Now, use Bayes' rule and the bound in (5.91) to write

$$\begin{aligned}&\Pr\{\cup_j(K_j < K/2)|\mathcal{S}_{\beta,N_{\max}}\} \\ &\qquad = \sum_{\mathcal{S}\in\mathcal{S}_{\beta,N_{\max}}} \Pr\{K_j < K/2|\mathcal{S}\}\Pr\{\mathcal{S}\} \\ &\qquad \leq \ (K/2)2^K\beta^{K/2},\end{aligned} \tag{5.92}$$

where in obtaining the inequality we used that for $\mathcal{S} \in \mathcal{S}_{\beta,N_{\max}}$ the decoding failure probability at $I_0^{(k^*)}$ satisfies $\Pr\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\} \leq \beta$ and that $\sum_{\mathcal{S}\in\mathcal{S}_{\beta,N_{\max}}} \Pr\{\mathcal{S}\} \leq 1$.

For $\beta = 1/8$ the latter bound reduces to $\Pr\{\cup_j(K_j < K/2)|\mathcal{S}_{\beta,N_{\max}}\} \leq (K/2)(1/2)^{K/2}$ which goes to zero as $K \to \infty$. Since $K \to \infty$ is implied when $J \to \infty$, we conclude from the latter that for any $\epsilon/(3N_{\max}) > 0$, $\exists\ J_0$ such that

$$\Pr\{K_j < K/2|\mathcal{S}_{\beta,N_{\max}}\} < \epsilon/(3N_{\max}), \tag{5.93}$$

for every $J > J_0$.

To bound the second term in (5.89), we invoke Lemma 6-[a] and Theorem 10. First, note that we can write

$$\begin{aligned}\Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}\} = &\Pr\{(N_B, N_A) > N_{\max}\} \\ &+ \Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}|(N_B, N_A) < N_{\max}\},\end{aligned} \tag{5.94}$$

Lemma 6-[a] guarantees that we can choose $N_{\max}$ sufficiently large so that

$$\Pr\{(N_B, N_A) > N_{\max}\} < \epsilon/3, \quad \forall J, \tag{5.95}$$

taking care of the the first term in (5.94). In the second term the numbers (N_B, N_A) of active-A and active-B users are given, and we can apply Theorem 10.

Note that since Theorem 10 is valid for any $k \leq K$, it must hold for $I_0^{(k^*)}$; and consequently, as $N_I := J - N_A - N_B > J - 2N_{\max} \to \infty$, we must have

$$\Pr\{I_0^{(k^*)} \notin \mathcal{C}_0|(N_B, N_A) < N_{\max}\} \to 0, \tag{5.96}$$

when the failure probability is *not* conditioned on $\mathcal{S}$.

Suppose that $\Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}|(N_B, N_A) < N_{\max}\} > \epsilon/3 \ \forall J$ and argue by contradiction. Indeed, if this were true we would have $\Pr\{I_0^{(k^*)} \notin \mathcal{C}_0|\mathcal{S}\} \geq \beta$ for a subset of network realizations $\{\overline{\mathcal{S}_{\beta,N_{\max}}}|(N_B, N_A) > N_{\max}\}$ with non-vanishing measure. But this is incompatible with (5.96) and consequently for any $\epsilon/3 > 0$, $\exists\ J > J_1$ such that

$$\Pr\{\overline{\mathcal{S}_{\beta,N_{\max}}}|(N_B, N_A) < N_{\max}\} < \epsilon/3. \tag{5.97}$$

Substituting (5.95) and (5.97) into (5.94), and the result of this operation along with (5.93) into (5.89), we finally obtain that

$$\Pr\{\cup_j (K_j < K/2)\} \leq \epsilon, \tag{5.98}$$

for arbitrary ϵ and all $J > \max(J_1, J_2)$. By definition, this implies the result in (5.63). □

Besides establishing our major claim previewed in Section 5.4.1, the asymptotic analysis of this section provides a series of byproduct remarks about OCRA:

Remark 19 *Average power constraint.* A consequence of the cooperators' clustering asserted by Theorem 10-[a] is that cooperation is limited to nearby idle users; and accordingly, the total transmitted power by any active communication is

$$\sum_{k=0}^{K_j} P(C_j^k) \approx (K_j+1)\frac{P_0}{K_j+1}\|B_j\|^\alpha/\xi = P_0\|B_j\|^\alpha/\xi. \tag{5.99}$$

Comparing (5.99) with rule [R2], we see that the average transmitted power in non-cooperative SSRA is equal to OCRA's phase-B power. The sole power increase is due to the phase-A power used to recruit cooperators, yielding the relation

$$P^{\text{OCRA}}(U_j) \approx (1+\rho)P^{\text{SSRA}}(U_j) \tag{5.100}$$

between the power required by OCRA and non-cooperative SSRA. Since $\rho \to 0$, we deduce that OCRA enables high order diversity with a small increase in average transmitted power.

Remark 20 *Maximum power constraint.* A maximum power constraint $P(U_j) \leq P_{\max}$ determines the AP's coverage area, since power control dictates that $\|U_j\|^\alpha \leq (\xi P_{\max}/P_0) := R_c^\alpha$. But since power in OCRA is contributed by K_j cooperators, we have

$$R_c^{\text{OCRA}} = (K_j)^{1/\alpha} R_c^{\text{SSRA}}. \tag{5.101}$$

This increase in coverage stems from the fact that users in OCRA transmit less power during more time.

Remark 21 *Network Area.* The proofs rely on the asymptotic behavior of the distance ratios in Lemma 5. This behavior does not depend on the radius of the network, implying that we can make it arbitrarily large. Accordingly, our major claims in Theorems 8 and 9 are valid for a fixed area network with increasing user density as well as for a fixed user density network with increasing area.

Remark 22 *OCRA with different physical layers.* It is known that diversity in wireless networks requires a transmitter that enables, a channel that provides, and a receiver that collects diversity. While results in this chapter have been derived for SSRA networks whose

suitability in enabling and collecting diversity is well appreciated, the advantage of OCRA is that it *generates* multipath diversity in a channel that originally did not provide it. This result depends on the spatial distribution of users and can be readily established for RA networks with different physical layers. The difference in these other cases will be the way in which the diversity is enabled and collected; but retaining the essential diversity-*providing* structure of a low power phase-A followed by a high order diversity phase-B will lead to claims analogous to Theorems 8 and 9.

5.6 Unslotted OCRA

Packet de-spreading at the AP is performed through multiplication with the appropriately delayed version of the spreading sequence $\mathbf{c}$. Indeed, multiplication by $\mathbf{c}(t - \tau_{C_j^k})$ allows the AP to recover the k^{th} copy of B_j's phase-B packet; and multiplication by $\mathbf{c}(t)$ allows idle users to detect $A_j's$ packet. Unfortunately, this requires knowledge of the delay $\tau_{C_j^k}$, and the only way of accomplishing this in RA is by having the AP check all the (virtually infinite) possible shifts τ. This complexity can be reduced by altering the PN selection rule to let the nodes choose a random shift at the beginning of time, communicate this selection to the AP and then use the same shift for the life of the network. A more elegant solution to this problem is through an unslotted protocol as we outlined for non-cooperative SSRA networks in [89].

In this unslotted version, active-A and active-B users choose a random time to start transmitting, but they spread their packets with an *unshifted* version of the common PN sequence. This entails replacing rules [S1]-[S2] and [S4]-[S5] with the following.

[U1] If U_j's queue is not empty, U_j enters phase-A with probability p and moves the first packet in the queue, $\mathbf{d}_{U_j} := \{d_{U_j}(l)\}_{l=0}^{L-1}$, to the phase-A buffer.

[U2] **Phase-A**: The transmission is as in [S2], but we include in the packet header the time T_{B_j} in which phase-B transmission is going to be attempted. The time T_{B_j} is chosen so that the transmission probability in each time unit is p.

[U4] U_j enters phase-B at time T_{B_j}.

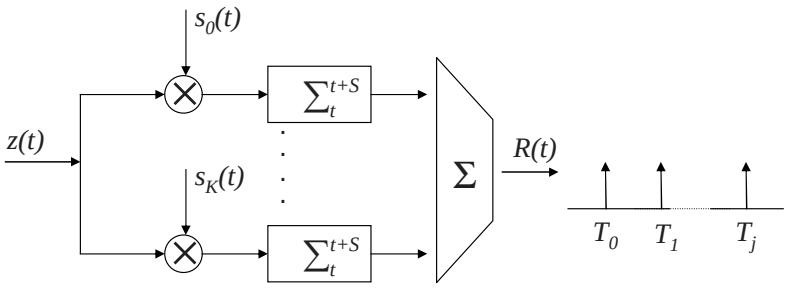

Figure 5.7: In unslotted OCRA, the correlator shown can be used to detect the starting times of a packet. Simulations corroborate that slotted and unslotted OCRA exhibit similar throughputs.

[U5] Phase-B: Transmission is as in [S5] but when spreading $\mathbf{d}_{U_j}$ the cooperator C_j^k uses the shift

$$\tau_{C_j^k} = \tau_k T, \tag{5.102}$$

with $\tau_0 = 0$ and $\tau_k \sim \mathcal{U}[0, \kappa - 1]$.

When expressed with respect to a common time reference, the equivalent of (5.48) for this unslotted system becomes

$$x_{A_j}(Sl + s) = \sqrt{P(A_j)}\; d_{A_j}(l) c(Sl + s - T_{A_j}) \sqcap (Sl + s - T_{A_j}) \tag{5.103}$$

where $\sqcap(t)$ is a unit-amplitude square pulse with nonzero support over $t \in (0, NL)$. Relying on (5.103), we can repeat the steps in Appendix A.2 to deduce that this spreading rule achieves statistical user separation at the idle users. Similarly, for the cooperative phase-B transmissions the counterpart of (5.49) is

$$\begin{aligned} x_{C_j^k}(Sl + s) = &\sqrt{P(C_j^k)}\; d_{B_j}(l) \\ &\times c(Sl + s - T_{B_j} - \tau_k T) \sqcap (Sl + s - T_{B_j}) \end{aligned} \tag{5.104}$$

with $k \in [0, K_j]$. Again, by following the steps in Appendix A.2 we can prove that this achieves statistical user separation at the AP.

The difference is that the first symbol in every packet is always spread by the same set of chips. Upon defining the (short) periodic sequences

$$c_k(t) = c(t - kSL \mod S), \qquad k \in [0, \kappa - 1]; \tag{5.105}$$

which amounts to periodically repeating the first S chips that spread the first symbol of any packet $\mathbf{x}_{C_j^k}$ or $\mathbf{x}_{A_j}$; the output of a continuous correlator matched to $s_k(t)$ can be used to detect the beginning of a packet; see also Fig 5.7. Indeed, the sum of the outputs of these correlators is

$$R(t) = \sum_{k=0}^{\kappa} \sum_{t'=t}^{t+S} c_k(t')z(t') = \sum_{k=0}^{\kappa} \sum_{t'=t}^{t+S} c(-kT)z(t'), \tag{5.106}$$

since we have that $c_k(t) = c(-kSL)$ in an interval of length S. But $\mathrm{E}(R(t)) = 0$, except when a packet started at time t, in which case $\mathrm{E}(R(t)) = \pm SP_0$, the sign being the value of the transmitted bit. Accordingly, the event $|R(t)| > SP_0/2$ can be used by the AP to identify the starting time of B_j's packet at $T_{B_j} = t$. A similar correlator with $\kappa = 0$ in (5.106) can be used by the idle users to identify the times T_{A_j}.

Thus, an unslotted version of OCRA reduces the challenging task of identifying the random shifts τ_{B_j} to the easier problem of identifying the random times T_{B_j}. Interestingly, the number of correlations computed does not change; what changes is that instead of taking κT correlations at the beginning of a slot, we take κ correlations during T times. The difference is, of course, that Theorems 8 and 9 (and all other results for that matter) apply to the unslotted version. In the next section, we simulate unslotted OCRA as defined by rules [S0], [U1]-[U2], [S3], [U4]-[U5] and [S6]-[S7] to unveil that as is usual in SSRA networks (see e.g., [43]) the throughput of this practically feasible unslotted version is accurately predicted by the theoretical results derived for the slotted version.

5.7 Simulations

We have established in this chapter that slotted OCRA operating over a Rayleigh fading channel can asymptotically achieve the throughput of an equivalent non-cooperative SSRA operating over an AWGN channel, promising an order of magnitude increase in throughput.

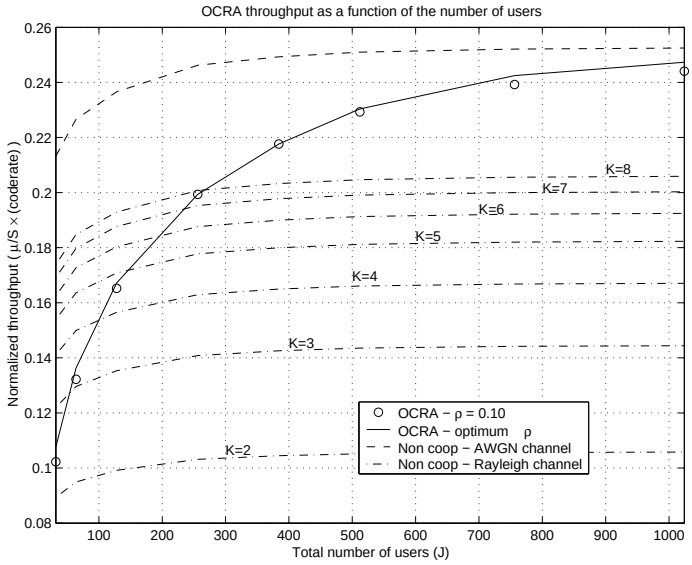

Figure 5.8: OCRA captures a significant part of the diversity advantage in mid-size networks; the MST for $J = 128$ is 2/3 the MST of SSRA over an AWGN channel ($\kappa = 10$, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

In this section, we explore three questions of significant practical importance that our theoretical results left only partially answered. These questions are: i) does slotted OCRA results carry over to unslotted OCRA? ii) how large the number of users should be to achieve a significant throughput increase? and iii) how do we select ρ and κ? To address i), we performed simulations for slotted and unslotted OCRA obtaining almost identical results in all the metrics studied; to avoid presenting virtually identical figures, we report only the figures pertaining to unslotted OCRA stressing the fact that they basically coincide with the curves for slotted OCRA. The answers to ii) and iii) are provided in the remainder of this section.

Consider first question ii) and refer to Fig. 5.8 where we depict unslotted OCRA's MST, $\eta_{\max}^{OCRA}$, as a function of the number of users J in a network with spreading gain $S = 32$, packet length $L = 1024$, and a 215/255 BCH code capable of correcting $t = 5$ errors used for FEC. A quick inspection of Fig. 5.8 reveals that convergence to AWGN throughput is rather slow since for J as large as 512 there is still a noticeable gap. Notwithstanding,

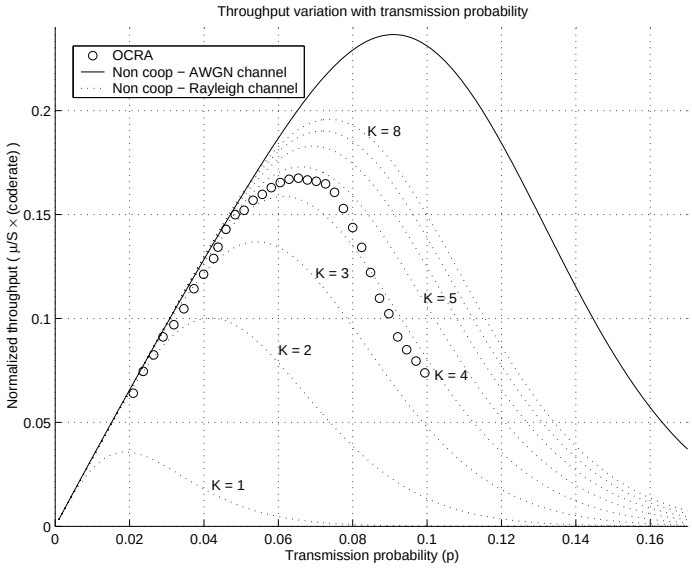

Figure 5.9: OCRA throughput with variable packet transmission probability p. In the range shown, OCRA's throughput remains between the throughput of non-cooperative SSRA over Rayleigh channels with diversity of order 4 and 5 ($\rho = 0.01$, $\kappa = 10$, $J = 128$, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

the throughput increase is rather fast; for $J = 64$ there is a threefold throughput increase ($\eta_{\max} = 0.04$ if the channel is Rayleigh), and for $J = 128$ OCRA's MST is 2/3 of the MST achieved by non-cooperative SSRA over an AWGN channel. Thus, while collecting the full diversity advantage requires an inordinately large number of users, OCRA can collect a significant percentage of it in moderate size networks, with a ratio $J/S \approx 4$. This behavior can be explained through the background curves that show the MST of non-cooperative systems with increasing diversity order. These curves illustrate the well understood behavior that the throughput increase when the diversity order goes from 2 to 3 is much larger than the increase when the diversity order goes from 7 to 8, [121]. Moreover, a large part of the potential increase is collected with order 5 diversity. As a diversity enabler, OCRA quickly achieves 5-order diversity when $J \approx 128$; but additional improvements in the diversity order translate to increasingly small throughput increments.

Similar conclusions can be drawn from the simulation with $J = 128$ users depicted in

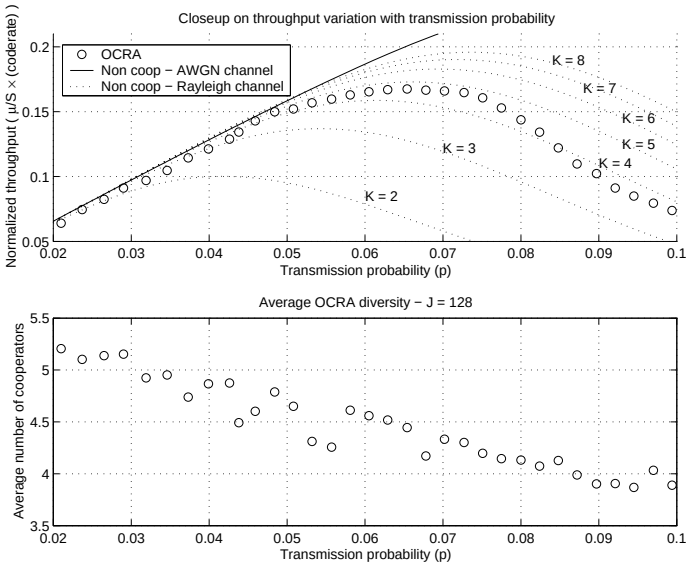

Figure 5.10: A closer look to Fig 5.9. OCRA's throughput is consistent with the fact that the average number of cooperators is between 4 and 5 ($\rho = 0.01$, $\kappa = 10$, $J = 128$, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

Figs. 5.9 and 5.10. For this case study, we show throughput and average diversity as a function of the transmission probability p. For the range of probabilities close to the MST, OCRA's throughput remains between the curves for 4 and 5-order diversity, consistent with the fact that the average degree of cooperation that users receive is between 4 and 5.

Turning our attention to question iii), let us recall the distinction between ρ-conditional MST in (5.61) and MST in (5.62). Interestingly, optimizing over (ρ, p) provides a small throughput increase with respect to optimizing over p only, as can be seen in Fig. 5.8. In this plot, the solid line depicts OCRA's MST and the circles depict the ρ-conditional MST, when we set $\rho = 0.01$. In the vast operational range shown, there is no noticeable difference between these two approaches. This has the important practical implication that we do not need to optimize ρ, removing a significant part of the added complexity that OCRA incurs relative to non-cooperative SSRA.

Finally, it is interesting to check our intuition about OCRA by looking at the network snapshots depicted in Figs. 5.11 and 5.12. OCRA effectively exploits wasted resources in

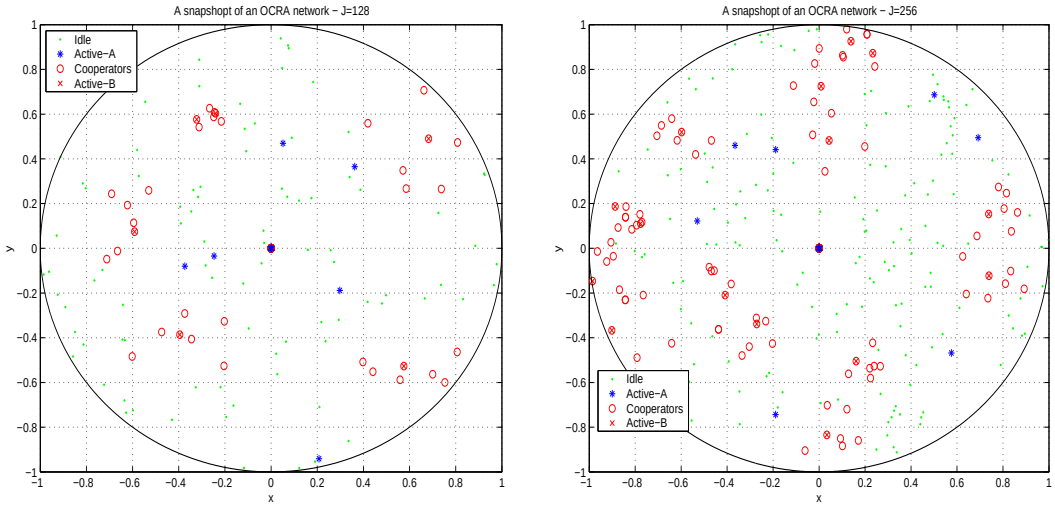

Figure 5.11: Snapshots of OCRA networks. OCRA effectively exploits the otherwise wasted cooperators' transmitters to provide user cooperation diversity ($p = p_{\max}(\rho)$ $\rho = 0.01$, $\kappa = 10$, $J = 128$ in left, $J = 256$ in right, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

non-cooperative RA, namely idle users' transmitters, as can be seen in Fig 5.11. In a conventional SSRA, only a small number of active-B users would be transmitting; whereas in OCRA, the cooperators are a significant percentage of the total number of users. This does not change as the number of users increases since when we go from $J = 128$, Fig. 5.11 (left) to $J = 256$, Fig. 5.11 (right), the number of cooperators per user increases so as to exploit the otherwise wasted cooperators' transmitters. It is also interesting to verify that as predicted by Theorem 10 the cooperators become clustered around the active-B user they are cooperating with.

The perspective of an active-A user can be summarized in the interference map depicted in Fig. 5.12. Each point in this map represents the total power received from all active-B users and their cooperators, and effectively represents the amount of noise in the active-A to idle users links. Thus, idle users in purple spots have low SINR and are not likely to be recruited as cooperators and idle users in green-yellow spots have large SINR and are likely to be recruited as cooperators. As the network size increases, the interference map

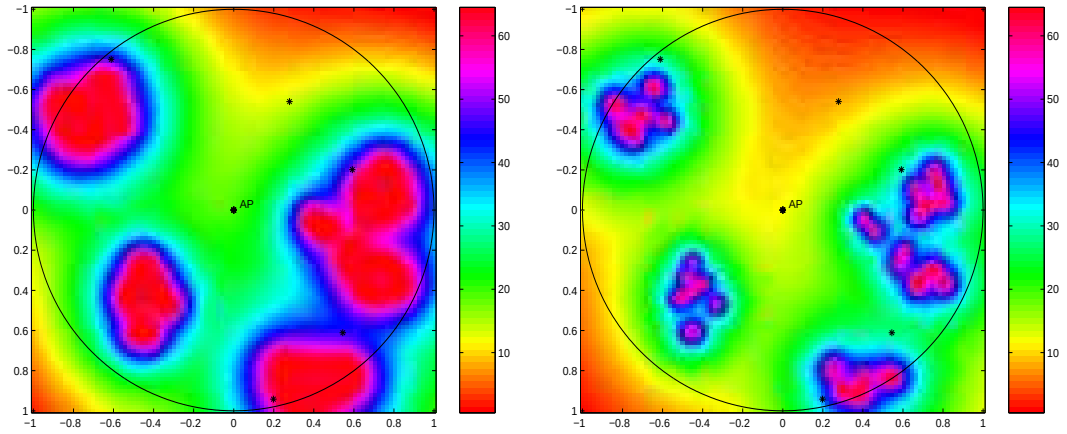

Figure 5.12: Interference maps. The color scale represents the total interference in dB received from active-B users at any point in space. As the number of users increases, the interference map remains essentially the same but the signal power received at idle users from active-A users increases. This translates in an increased number of idle users with good reception opportunities for active-A packets ($p = p_{\max}(\rho)$ $\rho = 0.01$, $\kappa = 10$, $J = 128$ in left, $J = 256$ in right, $S = 32$, $L = 1024$, 215/255 BCH code capable of correcting $t = 5$ errors).

is essentially unchanged by Lemma 6, but the signal power in the active-A to idle users links increases. This translates to an increase of the green-yellow area when the number of users increases from $J = 128$, Fig. 5.12 (left) to $J = 256$, Fig. 5.12 (right). Since users are uniformly distributed, this also translates to an increased number of idle users with good reception opportunities for active-A packets.

The simulations presented provide a reasonable answer to questions i) – iii) at the beginning of the section corroborating that: i) unslotted OCRA behaves as slotted OCRA; ii) the asymptotic behavior applies even to moderate-size networks having $J/S \approx 4$; and iii) $\rho \approx 0.1$ is a reasonable rule of thumb, and $\kappa \approx 10$ enables 4 to 6 diversity paths.

5.8 Summary

With the goal of migrating user cooperation benefits to random access channels, we introduced the OCRA protocol which we showed capable of effecting a significant throughput increase with respect to equivalent non-cooperative random access protocols. Testament to this significant advantage is the fact that as the number of users in the network increases, OCRA's throughput over Rayleigh fading links approaches that of the corresponding SSRA protocol over AWGN links, without an energy penalty. Accordingly, *OCRA has the capacity of rendering a wireless RA channel equivalent to a wireline one* from the throughput perspective. This is a striking difference with point to point cooperation, where the diversity comes at the expense of bandwidth expansion. The price paid is a modest increase in the complexity (and therefore cost) of the baseband circuitry.

Simulations demonstrated that our asymptotic results can be perceived in realistic-sized networks, since the asymptotic results manifest for moderate values of the total number of users.

The OCRA protocol relies on a two-phase transmission in which users first transmit with reduced power trying to reach nearby users, whose cooperation is thereby solicited for the subsequent slot. In this second slot, the (random) number of cooperators recruited transmit cooperatively to the destination. While a specific (spread spectrum) physical layer support was assumed, the same approach and results can be applied to other physical layers with the consequence of an intrinsic suitability of user cooperation as *the* form of diversity for random access networks.

5.9 Appendices

5.9.1 Other users' interference in OCRA 4

Signal reception at the AP

Substituting the explicit value of $z(Sl+s)$ in (5.50) into (5.53) and using the expression for the composite fading coefficient in (5.52) we can write the decision statistic $r_{C_{j_0}^{\kappa_0}}(l)$ as

$$r_{C_{j_0}^{\kappa_0}}(l)=h(C_{j_0}^{\kappa_0})h_n^*(C_{j_0}^{\kappa_0})d_{B_{j_0}}(l)+\sum_{\substack{j=1\\ j\neq j_0}}^{N_B}\sum_{k=0}^{K_j}\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;C_j^k)$$
$$+\sum_{\substack{k=0\\ \tau_k\neq\kappa_0}}^{K_{j_0}}\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;C_{j_0}^k)+\sum_{j=1}^{N_A}\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;A_j)+\tilde{n}(l) \tag{5.107}$$

where we used the notation (introduced after (5.11)) $\mathbb{I}(l;C_{j_0}^{k_0}\rightarrow AP;U)$ to represent the interference of user U to the aggregate link $C_{j_0}^{\kappa_0}\rightarrow AP$ for the transmission of the l^{th} bit. The first group of interference terms corresponds to the active-B users $B_j\neq B_{j_0}$, the second group to the cooperators of B_{j_0} that chose a different shift $\tau_k\neq\kappa_0$, and the third group to the active-A users. These interference terms are given by

$$\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;U)=\sqrt{P(U)}\;h(U)h_n^*(C_{j_0}^{\kappa_0})d_U(l)$$
$$\times\frac{1}{S}\sum_{s=0}^{S-1}c(Sl+s-\tau_U)c(Sl+s-\tau_{B_{j_0}}-\kappa_0T) \tag{5.108}$$

with U denoting alternatively C_j^k $j\neq j_0$, $C_{j_0}^k$ $\tau_k\neq\kappa_0$, and A_j.

Using the low autocorrelation property of long PN sequences, we obtain that if $\tau_{C_{j_0}^{\kappa_0}}\neq\tau_U$ – for what it suffices to have $\tau_{B_{j_0}}\neq\tau_{B_j}$, for $j\in[1,N_B]$, $j\neq j_0$– then

$$\mathrm{E}[\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;U)]=0 \tag{5.109}$$

$$\mathrm{var}[\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;U)]=(1/S)P(U)\mathrm{E}\left[|h(U)|^2\right] \tag{5.110}$$

$$\mathrm{E}[\mathbb{I}(l;C_{j_0}^{\kappa_0}\rightarrow AP;U_1)\mathbb{I}^*(l;C_{j_0}^{\kappa_0}\rightarrow AP;U_2)]=0. \tag{5.111}$$

Since when $\tau_{C_{j_0}^{\kappa_0}}\neq\tau_U$ all the random variables in (5.108) are independent, we have that: i) eq. (5.109) follows immediately since any of the involved random variables has zero

mean; ii) when computing the variance in (5.110) we have that $\mathrm{E}[h(U)h^*(U)] = \mathrm{E}[|h(U)|^2$, $\mathrm{E}[h_n^*(C_{j_0}^{\kappa_0})h_n^*(C_{j_0}^{\kappa_0})] = 1$, $\mathrm{E}[d_{U_j}(l)] = 1$, and among the S^2 cross-products involving the code $\mathbf{c}$ only S of them are not null; and iii) to establish (5.111) it suffices to note that $h(U_1)$ and $h(U_2)$ are independent and zero-mean.

Using property (5.109) we can see that none of the interference terms in (5.107) contributes to the mean of $r_{C_{j_0}^{\kappa_0}}(l)$ and consequently

$$\mathrm{E}[r_{C_{j_0}^{\kappa_0}}(l)] = \mathrm{E}\left[|h(C_{j_0}^{\kappa_0})|\right] = \sqrt{\frac{P_0 N(B_{j_0},\kappa_0)}{K_{j_0}+1}} d_{B_{j_0}}(l), \tag{5.112}$$

since the composite channel $h(C_{j_0}^{\kappa_0})$ contains $N(B_{j_0},\kappa_0)$ terms, each with power $P_0/(K_{j_0}+1)$. Likewise, (5.111) allows us to separate the variance in independent terms

$$\begin{aligned}\mathrm{var}[r_{C_{j_0}^{\kappa_0}}(l)] &= \mathrm{E}[\tilde{n}^2(l)] + \sum_{\substack{j=1\\ j\neq i}}^{N_B}\sum_{k=0}^{K_j} \mathrm{E}[|\mathbb{I}(l; C_{j_0}^{k_0}\!\rightarrow\! AP; C_j^k)|^2] \\ &+ \sum_{\substack{k=0\\ k\neq k_0}}^{K_j} \mathrm{E}[|\mathbb{I}(l; C_{j_0}^{k_0}\!\rightarrow\! AP; C_{j_0}^k)|^2] + \sum_{j=1}^{N_A} \mathrm{E}[|\mathbb{I}(l; C_{j_0}^{k_0}\!\rightarrow\! AP; A_j)|^2]\end{aligned} \tag{5.113}$$

Evaluating the expected values in (5.113) we obtain

$$\begin{aligned}\mathrm{var}[r_{C_{j_0}^{\kappa_0}}(l)] &= N_0 + (N_B-1)\frac{P_0}{S} \\ &\quad + [K_{j_0}+1-N(B_{j_0},\kappa_0)]\frac{P_0}{S(K_{j_0}+1)} + N_A\frac{\rho P_0}{S}\end{aligned} \tag{5.114}$$

where we used: i) property (5.110), ii) the power control rules $P(A_j)\mathrm{E}[|h(AP,A_j)|^2] = \rho P_0$ in (5.40) and $P(C_j^k)\mathrm{E}[|h(AP,C_j^k)|^2] = P_0/(K_j+1)$ in (5.41), and iii) that the number of summands in the second sum is $[K_{j_0}+1-N(B_{j_0},\kappa_0)]$,

From (5.112) and (5.114), the SINR in (5.54) follows from its definition.

Signal reception at idle users

Using once again the notation $\mathbb{I}(l; U_0\!\rightarrow\! I_i; U)$ to denote the interference of U to the communication of the l^{th} bit of the packet $\mathbf{d}_{U_0}$ from U_0 to I_i, the entries of the decision vector

in (5.56) can be written as

$$r_{U_0}(l) = \sqrt{P(U_0)}\; h(U_0, I_i) d_{U_0}(l) + \sum_{j=1}^{N_B} \sum_{k=0}^{K_j} \mathbb{I}(l; U_0 \rightarrow I_i; C_j^k) + \sum_{j=1}^{N_A - 1} \mathbb{I}(l; U_0 \rightarrow I_i; A_j) + \tilde{n}(l) \tag{5.115}$$

where $\mathrm{E}[P(U_0)|h(U_0, I_i)|^2] = P(U_0 \rightarrow I_i)$ is the power received from U_0 at I_i and is given by the pathloss model (5.3). The interference terms are given by [c.f. (5.55)]

$$\mathbb{I}(l; U_0 \rightarrow I_i; U) = \sqrt{P(U)}\; h(U, I_i) h_n^*(U_0, I_i) d_U(l) \times \frac{1}{S} \sum_{s=0}^{S-1} c(Sl + s - \tau_U) c(Sl + s) \tag{5.116}$$

where, as before, $\mathrm{E}[P(U)|h(U, I_i)|^2] = P(U_j \rightarrow I_i)$ can be obtained from (5.3).

The important observation is that for active-B transmissions, including active-B terminals and their cooperators, the autocorrelation property of PN codes yields that $\mathrm{E}[\mathbb{I}(l; U_0 \rightarrow I_i; U)] = 0$, $\mathrm{var}[\mathbb{I}(l; U_0 \rightarrow I_i; U)] = P(U \rightarrow I_i)/S$ and $\mathrm{E}[\mathbb{I}(l; U_0 \rightarrow I_i; U_1)\mathbb{I}^*(l; U_0 \rightarrow I_i; U_2)] = 0$ deterministically, since the 0^{th} PN shift is reserved for active-A users.

For active-A users however, the PN shifts are all equal and we have

$$\mathrm{E}[\mathbb{I}(l; U_0 \rightarrow I_i; A_j)] = 0, \tag{5.117}$$

$$\mathrm{var}[\mathbb{I}(l; U_0 \rightarrow I_i; A_j)] = P(A_j \rightarrow I_i), \tag{5.118}$$

$$\mathrm{E}[\mathbb{I}(l; U_0 \rightarrow I_i; A_{j_1})\mathbb{I}^*(l; U_0 \rightarrow I_i; A_{j_2})] = 0, \tag{5.119}$$

where (5.117) and (5.119) follow from the independence between different user's fading coefficients and the fact that in (5.118) the interfering power is not reduced by the spreading gain, as usual.

Using these properties, we can compute the expected value and the variance of $r_{U_0}(l)$; and from there, the SINR_0^i in (5.57).

5.9.2 Proof of Lemma 4

In order to have $\|U_j - U_0\| < r$, user U_j must lie in the region

$$U_j \in \mathcal{O}(0, R) \cap \mathcal{O}(U_0, r) := \mathcal{R}, \tag{5.120}$$

where $\mathcal{O}(o, r)$ denotes a circle with center o and radius r. The probability of U_j being in $\mathcal{R}$ is simply

$$F(r) = \frac{\text{area}(\mathcal{R})}{\pi R^2}. \tag{5.121}$$

The right inequality in (5.75) follows from (5.121) after noting that

$$\text{area}(\mathcal{R}) < \text{area}[\mathcal{O}(U_0, r)] = \pi r^2. \tag{5.122}$$

The left inequality in (5.75) requires considering the case in which the intersection of $\mathcal{O}(U_0, r)$ with $\mathcal{O}(0, R)$ subtracts most of the area from $\mathcal{O}(U_0, r)$. This happens when U_0 is at the border of $\mathcal{O}(U_0, r)$ and $r = 2R$. In this case,

$$\text{area}(\mathcal{R}) = \pi R^2 = \frac{\pi r^2}{4}. \tag{5.123}$$

QED. □

5.9.3 Proof of Lemma 5

The proofs for all events are similar. We prove the lemma for $e_2(N_I, \mathcal{K})$ that is the most representative, and sketch the proofs for the remaining events.

Remark 23 In the subsequent proofs we exploit the fact that active-A and active-B users' positions are independent. Indeed, users that enter phase-A in a given slot enter phase-B in the subsequent one regardless of whether they succeeded in recruiting cooperators or not. Furthermore, users enter phase-A regardless of their knowledge regarding the activity of neighboring nodes. This is rather "foolish" since we are allowing transmissions with small success probability, but nonetheless allowed to maintain independence between active-A and active-B users' positions. See also Remark 17.

Proof for event $e_2(N_I, \mathcal{K})$

To simplify notation define $\mathcal{K}' := \mathcal{K}/\rho^{1/\alpha}$. Recall that $F(r)$ is the distribution of $\|B_j - U_0\|$ given U_0, and note that since the positions of the N_B active-B users are assumed

independent, we have

$$\begin{aligned}\Pr\{\|B_0^{(1)} - U_0\| > r \ |U_0\} &= \Pr\left\{\bigcap_{j=1}^{N_B}(\|B_j - U_0\| > r)|U_0\right\} \\ &= (\Pr\{\|B_j - U_0\| > r \ |U_0\})^{N_B} \\ &= [1 - F(r)]^{N_B}. \end{aligned} \tag{5.124}$$

On the other hand, recall that $F(r)$ is also the CDF of $\|I_j - U_0\|$ and denote by $f_{I_0^{(K)}}(r)$ the pdf of $\|I_0^{(K)} - U_0\|$ given U_0. A basic result in order statistics is that [6, chap. 3]

$$f_{I_0^{(K)}}(r) = \frac{N_I!}{(K-1)!(N_I-K)!} F^{K-1}(r)[1 - F(r)]^{N_I-K} \frac{\partial F(r)}{\partial r}. \tag{5.125}$$

Applying Bayes' rule to the probability of $e_2(N_I, \mathcal{K})$ as given by (5.77) conditioned on U_0's position and using the expressions in (5.124) and (5.125), we obtain

$$\begin{aligned}&\Pr\{e_2(N_I, \mathcal{K}) \ |U_0\} = \\ &= \int_{-\infty}^{\infty} \Pr\left\{\|B_0^{(1)} - U_0\| > \mathcal{K}'r \ |I_0^{(K)} = r\right\} f_{I_0^{(K)}}(r) \ dr \\ &= \int_0^{r^*} [1-F(\mathcal{K}'r)]^{N_B} \frac{N_I!}{(K-1)!(N_I-K)!} \\ &\qquad \times F^{K-1}(r)[1-F(r)]^{N_I-K} \frac{\partial F(r)}{\partial r} \ dr \end{aligned} \tag{5.126}$$

where we also used that $B_0^{(1)}$ is independent of $I_0^{(K)}$, and we defined $r^* := \min\{r \text{ s.t. } F(\mathcal{K}'r) = 1\}$ that is the relevant upper limit of the integral, since the integrand is null for $r > r^*$.

Applying Lemma 4 to the distribution $F(r)$, we obtain the following inequality valid in $(0, r^*)$:

$$F(\mathcal{K}'r) \leq \frac{(\mathcal{K}'r)^2}{R^2} = 4\mathcal{K}'^2 \frac{r^2}{4R^2} \leq 4\mathcal{K}'^2 F(r), \tag{5.127}$$

which upon substituting in (5.126) and changing variables $u = F(r)$, yields

$$\begin{aligned}&\Pr\{e_2(N_I, \mathcal{K} \ |U_0)\} \geq \\ &\int_0^{1/4\mathcal{K}'^2} \left(1 - 4\mathcal{K}'^2 u\right)^{N_B} \frac{N_I!}{(K-1)!(N_I-K)!} u^{K-1}[1-u]^{N_I-K} du. \end{aligned} \tag{5.128}$$

We can expand the binomial $(1-4\mathcal{K}'^2)^{N_B}$ and interchange sum and integral to obtain

$$\begin{aligned}
&\Pr\{e_2(N_I,\mathcal{K}\,|U_0)\}\\
&\geq \sum_{l=0}^{N_B}(-1)^l\binom{N_B}{l}(2\mathcal{K}')^{2l}\int_0^{1/4\mathcal{K}'^2}\frac{N_I!\;u^{l+K-1}[1-u]^{N_I-K}}{(K-1)!(N_I-K)!}du\\
&:= \sum_{l=0}^{N_B}(-1)^l i_l,
\end{aligned} \tag{5.129}$$

where we defined i_l as the absolute value of the l^{th} summand of the previous expression.

All these integrals can be evaluated in closed form. In particular, i_0 is given by

$$\begin{aligned}
i_0 &:= \int_0^{1/4\mathcal{K}'^2}\frac{N_I!}{(K-1)!(N_I-K)!}u^{K-1}[1-u]^{N_I-K}\,du\\
&= \sum_{j=K}^{N_I}\binom{N_I}{j}(1/4\mathcal{K}'^2)^j\left(1-1/4\mathcal{K}'^2\right)^{N_I-j}.
\end{aligned} \tag{5.130}$$

The latter can be either computed directly or simply obtained by noting that the integral in (5.130) is the CDF of the K^{th} order statistic of a uniform random variable.

The summation in (5.130) can also be interpreted as the CDF of a binomial random variable with N_I trials and probability of success $\mathcal{K}'^{-2}/4$. As $N_I\to\infty$, the distribution converges to a normal and we have that

$$\begin{aligned}
\lim_{N_I\to\infty} i_0 &= \lim_{N_I\to\infty} Q\left(\frac{K-N_I/4\mathcal{K}'^2}{\sqrt{N_I}/2\mathcal{K}'}\right)\\
&= \lim_{N_I\to\infty} Q\left(\frac{K-\rho^{2/\alpha}N_I/4\mathcal{K}^2}{\rho^{1/\alpha}\sqrt{N_I}/2\mathcal{K}}\right)
\end{aligned} \tag{5.131}$$

where $Q(x) := \int_x^\infty 1/(\sqrt{2\pi})\exp(-u^2/2)du$ is the cumulative Gaussian function, and we used the definition of $\mathcal{K}'$ in the last equality. But note that if $K<\rho^{2/\alpha}N_I/4\mathcal{K}^2$, then the expression in (5.131) converges to 1, and this is true since the hypothesis $K/(\rho^{2/\alpha}N_I)\to 0$ implies that for any $4\mathcal{K}^2$ there exists a $K/(\rho^{2/\alpha}N_I)$ such that $(K/\rho^{2/\alpha}N_I)<1/4\mathcal{K}^2$. Accordingly, we established that

$$\lim_{N_I\to\infty} i_0 = 1. \tag{5.132}$$

Consider now the remaining integrals that can be bounded as follows:

$$\begin{aligned} i_l &:= \binom{N_B}{l} \left(2\mathcal{K}'\right)^{2l} \int_0^{1/4\mathcal{K}'^2} \frac{N_I!\; u^{l+K-1}[1-u]^{N_I-K}}{(K-1)!(N_I-K)!}\; du \\ &< \binom{N_B}{l} \left(2\mathcal{K}'\right)^{2l} \int_0^{1} \frac{N_I!\; u^{l+K-1}[1-u]^{N_I-K}}{(K-1)!(N_I-K)!}\; du \\ &= \binom{N_B}{l} (2\mathcal{K}')2l \frac{N_I!}{(K-1)!(l+K)\ldots(l+N_I)} \end{aligned} \tag{5.133}$$

where the inequality is obtained from the positivity of the integrand, and the second equality can be obtained after repeatedly integrating by parts. Moreover, it is easy to bound the factorials in the previous expression to obtain

$$i_l < \frac{1}{l!}\left(\frac{N_B\mathcal{K}'^2}{KN}\right)^l = \frac{1}{l!}\left(\frac{N_B\mathcal{K}^2}{K\rho^{2/\alpha}N_I}\right)^l. \tag{5.134}$$

But for $\rho^{2/\alpha}N_I/K \to \infty$ and $K \to \infty$, we have that $i_l \to 0$ for $l \neq 0$ for arbitrary $\mathcal{K}$. Taking limit in (5.129) and using the results summarized in (5.132) and (5.134), it follows that

$$\lim_{N_I\to\infty} \Pr\{e_2(N_I, \mathcal{K}\,|U_0)\} = 1. \tag{5.135}$$

To complete the proof, just note that (5.135) is a stronger result than the one desired, since the limit is conditioned on U_0. □

Proof for event $e_1(N_I, \mathcal{K})$

Note that if Lemma 4 is valid for all U_0, it is also valid unconditionally when averaged over all possible U_0's. From there, we obtain the inequality

$$\Pr\{\|U_0\| < r\} = \frac{r^2}{R^2} \leq 4\Pr\{\|B_j - U_0\| < r\}\;, \tag{5.136}$$

for arbitrary B_j. But now note that by definition $\|B_j - U_0\| \geq \|B_0^{(1)} - U_0\|$; and consequently,

$$\begin{aligned} \Pr\{\|U_0\| <& (\mathcal{K}/\rho)\|I_0^{(K)} - U_0\|\} \\ &\leq 4\Pr\{\|B_0^{(1)} - U_0\| < (\mathcal{K}/\rho)\|I_0^{(K)} - U_0\|\}. \end{aligned} \tag{5.137}$$

But the events involved in the previous inequality are the complements of $e_1(N_I, \mathcal{K})$ and $e_2(N_I, \mathcal{K})$, which implies that

$$1 - \Pr\{e_1(N_I, \mathcal{K})\} \leq 4[1 - \Pr\{e_1(N_I, \mathcal{K})]. \tag{5.138}$$

Since we just proved that $\Pr\{e_2(N_I, \mathcal{K})\} \to 1$, we deduce that $\Pr\{e_1(N_I, \mathcal{K})\} \to 1$. □

Proof for event $e_3(N_I, \mathcal{K})$

Repeat steps (5.124) to (5.135) in the proof for $e_2(N_I, \mathcal{K})$. □

5.9.4 Proof of Theorem 10

Let us first recall the following fact that will be used in the proof of claims [a] and [b].

Fact 1 If we have $\text{SINR}_0^k \to \infty$ in (5.57), then $\Pr\{I_k \in \mathcal{C}_0\} \to 1$. Indeed, if $\text{SINR}_0^k \to \infty$ then for all but a zero-measure set of fading channel realizations the packet transmitted by U_0 is correctly received by I_k. Likewise, if $\text{SINR}_0^k \to 0$ in (5.57), then $\Pr\{I_k \in \mathcal{C}_0\} \to 0$.

Proof of claim [a]

If $C_j^k \in \mathcal{C}_j$, then it successfully decoded B_j's active-A packet in the previous slot. Consider SINR_j^k for the reception of B_j's active-B packet by the user I_k in the previous slot that can be bounded by

$$\text{SINR}_0^k \leq \frac{P(U_o \to I_k)}{N_0} = \frac{\rho P_0}{N_0} \frac{\|B_j\|^\alpha}{\|B_j - I_j^{(k)}\|^\alpha}, \tag{5.139}$$

where we just considered the noise term and neglected the other users' interference.

Assuming that $\|B_j - I_k\|/\|B_j\| > \epsilon$ and letting $N_I \to \infty$ in (5.139), we obtain

$$\lim_{N_I \to \infty} \text{SINR}_0^k \leq \lim_{N_I \to \infty} \frac{\rho P_0}{\epsilon N_0} = 0. \tag{5.140}$$

But now recall Fact 1 to claim that since $\text{SINR}_0^k \to 0$ we must have

$$\lim_{N_I \to \infty} \Pr\{I_k \in \mathcal{C}_j\} = \lim_{N_I \to \infty} P_e^1(\text{SINR}_0^k) = 0\ . \tag{5.141}$$

Thus, if $\|B_j - I_k\|/\|B_j\| > \epsilon$ for some ϵ, then $I_k \notin \mathcal{C}_j$ with probability 1. It thus follows that for those that did become cooperators, (5.81) must hold true. In particular, it is true for $C_j^{(K_j)}$. □

Proof of claim [b]

We start by establishing a simple consequence of claim [a] in the following corollary:

Corollary 7 *The event*

$$e_4(N_I,\mathcal{K}) := \{\|U_0 - B_j\| > 2\|B_j - C_j^{(K_j)}\| \ \ \forall\ j = 1,\ldots,N_B\} \tag{5.142}$$

has probability 1 as the number of idle users $N_I \to \infty$; i.e.,

$$\lim_{N_I\to\infty} \Pr\{e_4(N_I,\mathcal{K})\} = 1\ . \tag{5.143}$$

Proof: Consider the complement of $e_4(N_I,\mathcal{K})$, and use the union bound and Lemma 5 to claim that

$$1 - \Pr\{e_4(N_I,\mathcal{K})\} < 4N_B \Pr\{\|B_j\| < 2\|B_j - C_j^{(K_j)}\|\}. \tag{5.144}$$

But the latter goes to 0 according to Theorem 10-[a], with $\epsilon = 1/2$.

We now continue with the proof of claim [b].

Proof - [b]: According to Fact 1 it suffices to prove that $\mathrm{SINR}_0^{(k)} \to \infty$ in probability, or equivalently,

$$\lim_{N_I\to\infty} \Pr\{\mathrm{SINR}_0^{(k)} > \mathcal{K}'\} = 1 \quad \forall\ \mathcal{K}' > 0\ . \tag{5.145}$$

The inverse SINR is given by (5.71) and can be rewritten as

$$\begin{aligned}(\mathrm{SINR}_0^{(k)})^{-1} = S^{-1}&\sum_{j=1}^{N_B}\sum_{i=0}^{K_j}\frac{P(C_{(j)}^i \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})} \\ &+ S^{-1}\sum_{j=1}^{N_A-1}\frac{P(A_0^{(j)} \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})} + \frac{N_0}{P(U_0 \to I_0^{(k)})}\end{aligned} \tag{5.146}$$

where we have just reordered the summands according to their closeness to U_0.

We will first bound the noise term. To this end, supposing that $e_1(N_I,\mathcal{K})$ is valid, we obtain

$$\begin{aligned}\frac{N_0}{P(U_0 \to I_0^{(k)})} = \frac{N_0}{\rho P_0}\frac{\|U_0 - I_0^{(k)}\|^\alpha}{\|U_0\|^\alpha} &< \frac{N_0}{\rho P_0}\frac{\|U_0 - I_0^{(K_1)}\|^\alpha}{\|U_0\|^\alpha} \\ &< \frac{N_0\rho^{\alpha-1}}{\mathcal{K}^\alpha P_0} < \frac{N_0}{\mathcal{K}^\alpha P_0},\end{aligned} \tag{5.147}$$

where the first inequality follows since $\|U_0 - I_0^{(k)}\| < \|U_0 - I_0^{(K_1)}\|$ holds by definition for $k \leq K_1$, and in the last inequality we used that $\rho < 1$ and $\alpha > 2$.

Consider now the active-B users' interference terms. Since the transmitted powers are proportional to the distance to the AP as per [h3], we have

$$\begin{aligned} P^{1/\alpha}(C_{(j)}^i \to I_0^{(k)}) &= \frac{(P_0/K_j)^{1/\alpha}\|C_{(j)}^i\|}{\|C_{(j)}^i - I_0^{(k)}\|} \\ &< \frac{P_0^{1/\alpha}}{K_j^{1/\alpha}} \frac{\|I_0^{(k)}\| + \|C_{(j)}^i - I_0^{(k)}\|}{\|C_{(j)}^i - I_0^{(k)}\|} \\ &= \frac{P_0^{1/\alpha}}{K_j^{1/\alpha}} \left[1 + \frac{\|I_0^{(k)}\|}{\|C_{(j)}^i - I_0^{(k)}\|}\right]. \end{aligned} \tag{5.148}$$

where the inequality follows from the triangle inequality applied to the triangle with vertices $AP, I_0^{(k)}, C_{(j)}^i$. Application of the same inequality to the triangle $AP, U_0, I_0^{(k)}$, yields

$$\|I_0^{(k)}\| < \qquad \|U_0\| + \|U_0 - I_0^{(k)}\| < \|U_0\| + \|U_0 - I_0^{(K)}\|, \tag{5.149}$$

where the second inequality follows from the definition of $I_0^{(k)}$ (the k^{th} closest to U_0 idle user), and the fact that $k \leq K$. Applying once again the triangle inequality to the triangles $I_0^{(k)}, B_0^{(j)}, C_{(j)}^i$ and $U_0, B_0^{(j)}, I_0^{(k)}$, yields (see also Fig. 5.13)

$$\begin{aligned} \|C_{(j)}^i - I_{(0)}^{(k)}\| &> \|U_0 - B_0^{(j)}\| - \|U_0 - I_0^{(k)}\| - \|B_0^{(j)} - C_{(j)}^i\| \\ &> \|U_0 - B_0^{(j)}\| - \|U_0 - I_0^{(K)}\| - \|B_0^{(j)} - C_{(j)}^{(K_j)}\| \\ &> 1/2\|U_0 - B_0^{(j)}\| - \|U_0 - I_0^{(K)}\| \\ &> 1/2\|U_0 - B_0^{(1)}\| - \|U_0 - I_0^{(K)}\|. \end{aligned} \tag{5.150}$$

In deriving the second inequality we used that $\|U_0 - I_0^{(k)}\| < \|U_0 - I_0^{(K)}\|$ and $\|B_0^{(j)} - C_{(j)}^i\| < \|B_0^{(j)} - C_{(j)}^{(K_j)}\|$ which follows by definition since $k \leq K$ and $i \leq K_j$. In the third inequality, we assumed the validity of $e_4(N_I, \mathcal{K})$; and the fourth one follows from $\|U_0 - B_0^{(j)}\| > \|U_0 - B_0^{(1)}\|$, which also is valid by definition.

If we also assume that the event $e_2(N_I, \mathcal{K})$ holds, we obtain that [c.f., (5.77), (5.150)]

$$\|C_{(j)}^i - I_0^{(k)}\| > (\mathcal{K}/2\rho^{1/\alpha} - 1)\|U_0 - I_0^{(K)}\|. \tag{5.151}$$

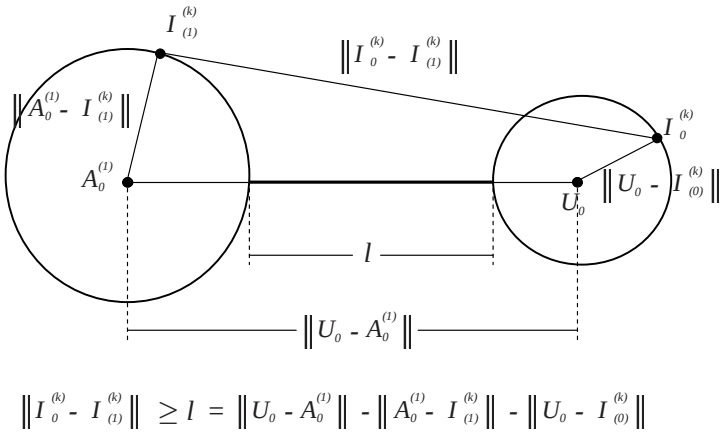

Figure 5.13: Repeated use of the triangle inequality bounds the SNR with the distance quotients considered in Lemma 5.

And the interfering power received at $I_0^{(k)}$ from $C_{(j)}^j$ can be bounded as [c.f., (5.148), (5.151)]

$$P^{1/\alpha}(C_{(j)}^i \to I_0^{(k)}) \;<\; \frac{P_0^{1/\alpha}}{K_j^{1/\alpha}} \times \left[1 + \frac{\|U_0\|}{(\mathcal{K}/2\rho^{1/\alpha} - 1)\|U_0 - I_0^{(K)}\|} + \frac{1}{\mathcal{K}/2\rho^{1/\alpha} - 1}\right]. \tag{5.152}$$

On the other hand, the power received at $I_0^{(k)}$ from U_0 is $P^{1/\alpha}(U_0 \to I_0^{(k)}) = (\rho P_0)^{1/\alpha}\|U_0\|$ / $\|U_0 - I_0^{(k)}\| > (\rho P_0)^{1/\alpha}\|U_0\|/\|U_0 - I_0^{(K)}\|$, from where we arrive at

$$\left[\frac{P(C_{(j)}^i \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})}\right]^{1/\alpha} < \frac{1}{(\rho K_j)^{1/\alpha}} \left[\frac{1}{\mathcal{K}/2\rho^{1/\alpha} - 1} + \left(1 + \frac{1}{\mathcal{K}/2\rho^{1/\alpha} - 1}\right) \frac{\|U_0 - I_0^{(K)}\|}{\|U_0\|}\right]. \tag{5.153}$$

Finally, note that if we assume that $e_1(N_I, \mathcal{K})$ is also true, we obtain the bound [c.f., (5.77), and (5.153)]

$$\frac{P(C_{(j)}^i \to I_0^{(k)})}{P(U_0 \to I_0^{(k)})} < \frac{1}{K_j \mathcal{K}^\alpha} f_B^\alpha(\mathcal{K}), \tag{5.154}$$

with $f_B(\mathcal{K})$ being a bounded function, since it is continuous and $\lim_{\mathcal{K}\to\infty} f_B(\mathcal{K}) = 4$.

Consider finally the active-A users' interference term that can be bounded by repeating the steps in (5.148) - (5.154), but instead of assuming the validity of the events $e_2(N_I, \mathcal{K})$ and $e_4(N_I, \mathcal{K})$ to go from (5.150) to (5.151), we assume that $e_3(N_I, \mathcal{K})$ is true. These steps yield

$$\left[\frac{P(A_0^{(j)} \to I_0^{(k)}}{P(U_0 \to I_0^{(k)})}\right]^{1/\alpha} < \left[\frac{1}{\mathcal{K}-1} + \left(\frac{\mathcal{K}}{\mathcal{K}-1}\right)\frac{\|U_0 - I_0^{(K)}\|}{\|U_0\|}\right] \tag{5.155}$$

from where the assumed validity of $e_1(N_I, \mathcal{K})$ leads to [c.f. (5.77) and (5.155)]

$$\left[\frac{P(A_0^{(j)} \to I_0^{(k)}}{P(U_0 \to I_0^{(k)})}\right] < \frac{1}{\mathcal{K}^\alpha} f_A^\alpha(\mathcal{K}), \tag{5.156}$$

with $f_A(\mathcal{K})$ bounded for the same reasons $f_B(\mathcal{K})$ is.

We can now combine the bounds in (5.147), (5.154) and (5.156) and the convexity of potential functions, $g(x) = x^\alpha$, with $\alpha > 1$, to conclude that if the events $\{e_l(N_I, \mathcal{K})\}_{l=1}^4$ hold true, then

$$\begin{aligned}(\mathrm{SINR}_0^{(k)})^{-1} &< N_B \frac{f_B^\alpha(\mathcal{K})}{S\mathcal{K}^\alpha} + (N_A - 1)\frac{f_A^\alpha(\mathcal{K})}{S\mathcal{K}^\alpha} + \frac{N_0}{\mathcal{K}^\alpha P_0} \\ &= \frac{1}{\mathcal{K}^\alpha}\left[(N_A - 1) f_B(\mathcal{K})/S + N_B f_A(\mathcal{K})/S + \frac{N_0}{P_0}\right] \\ &< \frac{\zeta}{\mathcal{K}^\alpha},\end{aligned} \tag{5.157}$$

for some constant ζ. Consequently, the probability that (5.157) is satisfied is larger than the probability of all four $\{e_l(N_I, \mathcal{K})\}_{l=1}^4$ holding true, and thus

$$\Pr\left\{(\mathrm{SINR}_0^{(k)})^{-1} < C/\mathcal{K}^\alpha\right\} > \Pr\left\{\bigcap_{l=1}^4 e_l(N_I, \mathcal{K})\right\}. \tag{5.158}$$

To complete the proof, apply the union bound to the intersection in (5.158) to obtain

$$\Pr\left\{(\mathrm{SINR}_0^{(k)})^{-1} < C/\mathcal{K}^\alpha\right\} > 1 - \sum_{l=1}^4 (1 - \Pr\{e_l(N_I, \mathcal{K}\}). \tag{5.159}$$

But according to Lemma 5, the four probabilities considered converge to 1 as $N_I \to \infty$, and we obtain that

$$\lim_{N_I\to\infty} \Pr\left\{(\mathrm{SINR}_0^{(k)}) > \mathcal{K}'\right\} = 1, \tag{5.160}$$

with $\mathcal{K}' := \mathcal{K}^\alpha/\zeta$. But as noted before, (5.160) implies that the PEP converges to 0, and (5.82) follows readily. □

Chapter 6

Future work

Shifting the routing paradigm from finding shortest paths in a graph to solving convex optimization problems as we discussed in Chapters 2 and 3 opens up the possibility to address a plethora of novel routing problems. Indeed, many rate maximizing criteria of practical interest lead to simple convex optimization problems. We contend that this fact, besides its intrinsic value, enables solution of additional routing problems that have been deemed intractable. We discuss some future problems in the next sections.

6.1 Robust optimal routing

While there is implicit robustness built into the stochastic routing protocols (SRP), of Chapters 2 and 3, a formulation that *optimizes* resilience against link fades and/or intentional attacks is certainly of interest. Say that $\mathbf{R}_0$ is the observed reliability matrix but due to link fades and/or intentional attacks the actual matrix $\mathbf{R}$ deviates from $\mathbf{R}_0$. We can capture this effect by modeling $\mathbf{R} \in \mathcal{R}$, where $\mathcal{R}$ is a set containing possible realizations of $\mathbf{R}$, e.g., all matrices including up to a 20% degradation in every R_{ij} entry. A robust routing formulation maximizes the rate $\boldsymbol{\rho}$ subject to the constraint that the optimal routing matrix $(\mathbf{T}^*, \mathbf{K}^*)$ is feasible for any $\mathbf{R} \in \mathcal{R}$; i.e.,

$$(\mathbf{K}^*, \mathbf{T}^*) = \arg \max_{\mathbf{K} \in \mathcal{K}, \mathbf{R} \in \mathcal{R}} \; f[(\mathbf{I} - \mathbf{K}_0)\mathbf{1}]. \tag{6.1}$$

As long as $\mathcal{R}$ is convex this problem also is. If we further require $R_{ij} \in [R_{ij}^{\min}, R_{ij}^{\max}]$, then (6.1) is a linear program.

A different approach to optimizing robustness is to consider that each entry R_{ij} of $\mathbf{R}$ is a random variable with known mean $\bar{R}_{ij}$ and variance Σ_{ij}. In this case, the resulting rates ρ_j are also random variables with mean $\bar{\rho}_j$ and variance σ_j. Two formulations of interest in this setup are to: i) maximize the average rate subject to a maximum tolerable variance, i.e., $\max_{\sigma \leq \sigma^{\min}} f(\bar{\rho})$; and ii) minimize the variance subject to a minimum acceptable rate, i.e., $\min_{\bar{\boldsymbol{\rho}} \geq \bar{\boldsymbol{\rho}}^{\min}} g(\sigma)$.

6.2 Routing in ad-hoc networks

In an ad-hoc network every terminal is a potential destination. Mimicking notation in Chapter 2, let $\{\boldsymbol{\rho}^{(j)}\}_{j=1}^{J}$ denote the arrival rates for delivery to node U_j. With $\boldsymbol{\lambda}^{(j)}$ denoting the corresponding departure rates and $\mathbf{T}^{(j)}, \mathbf{K}^{(j)}$ the routing matrices, we deduce that $\boldsymbol{\rho}^{(j)} = (\mathbf{I} - \mathbf{K}^{(j)})\boldsymbol{\lambda}^{(j)}$ [cf. (2.24), with $\mathbf{K} = \mathbf{K}_0$]. Interestingly, each matrix $\mathbf{K}^{(j)} \in \mathcal{K}$ adheres to the same set of constraints considered in (3.1). The difference is that there are now many outgoing flows implying that the constraint $\mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}$ is replaced by $\mathbf{0} \preceq \sum_{j=1}^{J} \boldsymbol{\lambda}^{(j)} \preceq \mathbf{1}$. Rate maximizing routes in ad-hoc networks can thus be pursued by solving

$$\begin{aligned}(\mathbf{K}^*, \mathbf{T}^*) = \arg\max \;\; & f[(\mathbf{I} - \mathbf{K}^{(1)})\boldsymbol{\lambda}^{(1)}, \ldots, (\mathbf{I} - \mathbf{K}^{(J)})\boldsymbol{\lambda}^{(J)}] \\ \text{s.t.} \;\; & \mathbf{K}^{(j)} \in \mathcal{K}, \;\; \mathbf{0} \preceq \sum_{j=1}^{J} \boldsymbol{\lambda}^{(j)} \preceq \mathbf{1}. \end{aligned} \tag{6.2}$$

As formulated here, stochastic routing in ad-hoc networks leads to a bilinear program. We will pursue reformulations of (6.2) analogous to those in Chapter 2.

6.3 Cross-layer optimization

As described in Chapter 2, matrix $\mathbf{R}$ is chiefly determined by transmitted power. If terminals transmit over orthogonal channels, then $R_{ij}(P_j)$ and the specific functional dependence changes with, e.g., the fading model. In contention- or interference-limited networks we have

that in general $\mathbf{R}(\mathbf{p}, \boldsymbol{\lambda})$, even though we can in many cases use $\mathbf{R}(\mathbf{p}, \mathbf{1})$ as an upper bound on achievable rates. In a cross-layer optimal formulation it will be interesting to jointly optimize the routing matrix and the vector $\mathbf{p}$ of transmission powers to obtain

$$(\mathbf{K}^*, \mathbf{T}^*, \mathbf{p}^*) = \arg\max \ f[(\mathbf{I}-\mathbf{K})\mathbf{1}] \\ \text{s.t.} \ \ K_{ij} = R_{ij}(\mathbf{p}), \ \ \mathbf{K}^T\mathbf{1} = \mathbf{1}, \ \ \mathbf{T}^T\mathbf{1} = \mathbf{1}, \ \ \mathbf{0} \preceq \mathbf{p} \preceq \mathbf{p}^{(\max)}. \quad (6.3)$$

Depending on $R_{ij}(P_j)$, this problem can be tractable or not. Our preliminary analysis suggests that for block fading channels and orthogonal transmissions at sufficiently high SNR, (6.3) belongs to a class of convex optimization problems called geometric programs. For interference-limited networks (6.3) amounts to a non-convex signomial program.

6.4 Opportunistic routing

The SRPs in Chapters 2 and 3 do not fully exploit the broadcast nature of the wireless channel. Indeed, before transmission U_j tags the packet with its intended destination, say U_i, thus preventing the possibility of terminals $U_k \neq U_i$ keeping a successfully decoded packet. An alternative strategy could be for terminals $U_i \neq U_j$ to take independent decisions as to whether they keep a successfully decoded packet or not. This opportunistic approach can be captured in our framework by simply requiring $K_{ij} \leq R_{ij}$ for $i \neq j$. The packet remains in U_j's queue if it is not kept by any terminal, i.e., $K_{jj} = \prod_{i=1}^{J+J_{\text{ap}}}(1-K_{ij})$, where we supposed that terminals make independent decisions on whether to keep correctly decoded packets. In short, a worthwhile future direction is to find opportunistic routes as

$$\mathbf{K}^* = \arg\max \ f[(\mathbf{I}-\mathbf{K})\boldsymbol{\lambda}] \\ \text{s.t.} \ \ K_{ij} \leq R_{ij}, \ \ \prod_{i=1}^{J+J_{\text{ap}}}(1-K_{ij}) = K_{jj}, \ \ \mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}. \quad (6.4)$$

Interestingly, this is also a signomial program suggesting that techniques developed to solve cross-layer optimization routing problems can be used to solve opportunistic routing problems as well. Note that in (6.4) we are allowing packet duplication. We foresee that for the same $\mathbf{R}$ the stability region for opportunistic routing is larger than the non-opportunistic region in (2.29).

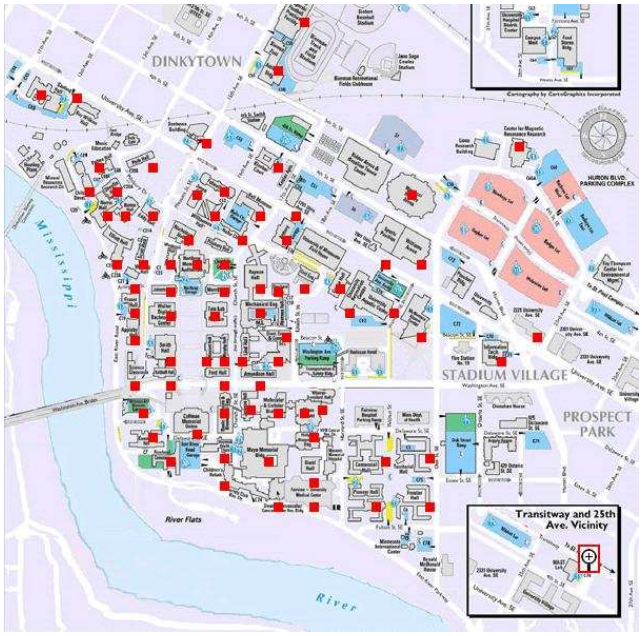

Figure 6.1: More than 400 wireless access points provide seamless 802.11 coverage throughout the UoM Twin cities campus.

6.5 Experiments and trials

The University of Minnesota (UoM) Twin cities campus is the second largest university campus in the nation boasting a student population of 51,175, thriving in a 2,730-acre urban setting [66]. To serve this student population, as well as faculty and staff, the Office of Information Technology (OIT) has deployed a 802.11b ("Wi-Fi") wireless network with 421 access points (APs) providing seamless coverage throughout campus; see also Fig. 6.1. Even though the 802.11b specification provides 11 channels, they overlap so that at most three channels can be used in the same space. In the UoM Wi-Fi network, channels 1, 6, and 11 are being used. Channel 11 is reserved for the campus-wide infrastructure deployed by OIT, whereas channel 1 is reserved for departmentally deployed units and channel 6 for future uses such as adding additional capacity or filling in weak spots. Most of the APs on campus work on the OIT-operated channel 11 with channel 1 being pervasive in technology-oriented departments and a few APs operating in channel 6 to cover difficult spots around campus. Its large size, heavy traffic, large number of users, and the inherent

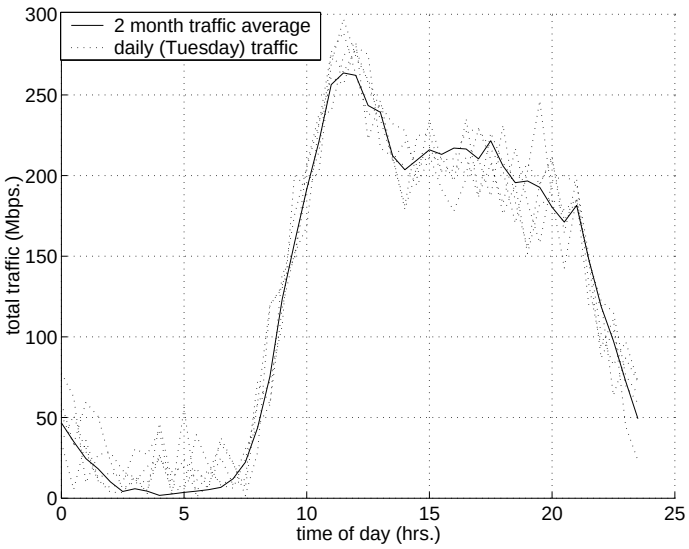

Figure 6.2: The UoM wireless network operates far from peak capacity except during the late morning to early afternoon rush-hour, leaving significant spare capacity for research trials.

wireless propagation difficulties of a urban setting make the UoM 802.11b network an ideal test-bed for the protocols and algorithms proposed in this thesis.

Consistent with the research initiatives of the UoM strategic positioning process [66] it is possible to test our protocols and algorithms using the UoM 802.11b network. We have gathered data to characterize the daily traffic behavior leading to the results summarized in Fig. 6.2. We show there how traffic (in Mbps) varies during different times of the day, plotting several individual data sequences as well as the average of 9 different days. As expected, traffic is negligible in the evenings and increases during the morning to reach its maximum during the late-morning to early-afternoon rush-hour; we then have a small decrease but traffic remains steady for the rest of the afternoon to finally start decreasing as the night approaches.

Interestingly, we also observed weekly variations with a characteristic pattern for Mondays, a different one for Tuesdays and so on (data in Fig. 6.2 is for all Tuesdays during September and October of 2005).

Taking into account the daily traffic distribution and the 802.11b channel assignment

we have developed an experiment and trial protocol consisting of 4 levels of experiments classified according to the risk of service disruptions:

[**L1**] **Data collection.** Passive data collection *not* entailing a traffic increase can be carried at any time during the day. Active data collection entailing traffic increase can be done between 8 pm and 10 am of the following day but never occupying more than 10% of installed capacity.

[**L2**] **Invasive experiments.** Experiments entailing traffic increases of more than 10% of installed capacity and/or requiring reconfiguration of APs will be performed from 11 pm to 7 am.

[**L3**] **Friendly users' trial.** System trials involving software installed in terminals of willing end users can be done in the reserved channel 6. These trials cannot use OIT-operated APs.

[**L4**] **Trial.** System trials in channel 11, using OIT-operated APs can be done after successful completion of [L3].

According to this plan the path conducing to the development of a working SRP should start with an [L1] experiment to collect information regarding the packet success probability matrix $\mathbf{R}$, i.e., its rate of change to understand how accurate the $\mathbf{R}$ estimate can be. We can then follow up with an [L2] experiment in which we reconfigure APs to communicate with each other and compare our SPRs with traditional routing alternatives to verify that the gains predicted in theory actually materialize in practice. After successful completion of this stage it is possible to move on to a friendly user trial with the Electrical and Computer Engineering departmental wireless network (level [L3]), and finally to a trial open to the whole UoM community (level [L4]).

Bibliography

[1] "EIA/TIA Interim Standard, Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard," *Qualcomm Inc.*, April 1992.

[2] N. Abramson, "Multiple access in wireless digital networks," vol. 82, pp. 1360–1370, September 1994.

[3] D. Aguayo, J. Bicket, S. Biswas, G. Judd, and R. Morris, "Link-level measurements from an 802.11b mesh network," *ACM SIGCOMM Computer Commun. Review*, vol. 34, pp. 121–132, October 2004.

[4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 1451–1458, October 1998.

[5] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," *IEEE Trans. Inf. Theory*, 2006 (to appear).

[6] N. Balakrishnan and A. C. Cohen, *Order Statistics and Inference Estimation Methods.* Academic Press Inc., 1991.

[7] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Process. Mag.*, vol. 21, pp. 59–68, September 2004.

[8] D. Bertsekas and R. Gallager, *Data networks.* Prentice-Hall, Inc., 2nd ed., 1992.

[9] D. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Athena Scientific, 2nd ed., 1997.

[10] S. Biswas and R. Morris, "ExOR: opportunistic multi-hop routing for wireless networks," *ACM SIGCOMM Computer Commun. Review*, vol. 35, pp. 133–144, October 2005.

[11] P. Bose, P. Morin, I. Stojmenovic, and J. Urrutia, "Routing with guaranteed delivery in ad hoc wireless networks," *Wireless Netw.*, vol. 7, pp. 609–616, November 2001.

[12] G. E. P. Box, *Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson eds. Robustness in statistics*. Academic Press, 1979.

[13] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[14] X. Cai, Y. Yao, and G. B. Giannakis, "Achievable rates in low-power relay-links over fading channels," *IEEE Trans. Commun.*, vol. 53, pp. 184–194, Jan. 2005.

[15] A. Cano-Pleite, T. Wang, and G. B. Giannakis

[16] A. Cano-Pleite, T. Wang, A. Ribeiro, and G. B. Giannakis, "Link-adaptive distributed coding for multi-source cooperation," in *Proc. Global Telecommun. Conf.*, San Francisco, CA, November 27 - December 1, 2006 (to appear). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[17] D. Chen and J. N. Laneman, "Modulation and demodulation for cooperative diversity in wireless systems," *IEEE Trans. Wireless Commun.*, 2006 (to appear).

[18] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 104–116, January 2005.

[19] M. Chiani, A. Conti, and V. Tralli, "Further results on convolutional code search for block-fading channels," *IEEE Trans. Inf. Theory*, vol. 50 , no. 6, pp. 1312–1318, June 2004.

[20] T. M. Cover and A. A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, pp. 572–584, September 1979.

[21] DARPA, "The next generation (XG) program," http://www.darpa.mil/ato/programs/xg/index.htm.

[22] D. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," in *Proc. Int. ACM Conf. Mobile Computing, Networking*, pp. 134–146, San Diego, CA, September 14-19, 2003.

[23] G. Dimic, N. D. Sidiropoulos, and R. Zhang, "Signal processing and queuing tools for MAC-PHY cross-layer design," *IEEE Signal Process. Mag.*, vol. 21, pp. 40–50, September 2004.

[24] E. Dinan and B. Jabbari, "Spreading codes for direct sequence CDMA and wideband CDMA cellular networks," *IEEE Commun. Mag.*, vol. 36, pp. 48–54, September 1998.

[25] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," in *Proc. of ACM SIGCOMM*, pp. 133–144, Portland, OR, August 30 - September 3, 2004.

[26] T. Eng and L. B. Milstein, "Coherent DS-CDMA performance in nakagami multipath fading," *IEEE Trans. Commun.*, vol. 43, pp. 1134–1143, February 1995.

[27] A. Ephremides, "Energy concerns in wireless networks," *IEEE Wireless Commun.*, vol. 9, pp. 48–59, August 2002.

[28] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and MAC for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 1514–1524, August 2006.

[29] Federal Communications Commission, *Et docket no. 03-322, notice of proposed rule making and order.* December 2003.

[30] D. Ganesan, R. Govindan, S. Shenker, and D. Estrin, "Highly-resilient, energy-efficient multipath routing in wireless sensor networks," *ACM SIGMOBILE Mobile Computing and Commun. Review*, vol. 5, pp. 11–25, October 2001.

[31] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong (eds.), *Signal processing advances in wireless and mobile communications - Volume I, trends in channel estimation and equalization.* Prentice-Hall, September 2000.

[32] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, "On the capacity of a cellular CDMA system," vol. 40, pp. 303–312, May 1991.

[33] A. J. Goldsmith and S. B. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Commun.*, vol. 9, pp. 8–27, August 2002.

[34] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, March 2002.

[35] M. Haenggi, "On routing in random Rayleigh fading networks," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1553–1562, July 2005.

[36] M. Haenggi and D. Puccinelli, "Routing in ad hoc networks: a case for long hops," *IEEE Commun. Mag.*, vol. 43, pp. 93–101, October 2005.

[37] B. Hassibi and H. Vikalo, "On the expected complexity of sphere decoding," in *Proc. Asilomar Conf. on Signals, Systems, Computers.*

[38] G. Holland, N. Vaidya, and P. Bahl, "A rate-adaptive MAC protocol for multi-hop wireless networks," in *Proc. Int. ACM Conf. Mobile Computing, Networking*, Rome, Italy.

[39] IEEE, "IEEE standard for information technology- telecommunications and information exchange between systems-local and metropolitan area networks- specific requirements Part II: wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11g-2003 (Amendment to IEEE Std 802.11,*

1999 Edn. (Reaff 2003) as amended by IEEE Stds 802.11a-1999, 802.11b-1999, 802.11b-1999/Cor 1-2001, and 802.11d-2001), 2003.

[40] M. Janani, A. Hedayat, T. Hunter, and A. Nosratinia, "Coded cooperation in wireless communications: space-time transmission and iterative decoding," *IEEE Trans. Signal Process.*, vol. 52, pp. 362 – 371, February 2004.

[41] H. Jiang, W. Zhuang, and X. Shen, "Cross-layer design for resource allocation in 3G wireless networks and beyond," *IEEE Commun. Mag.*, vol. 43, pp. 120–126, December 2005.

[42] B. Johansson, P. Soldati, and M. Johansson, "Mathematical decomposition techniques for distributed cross-layer optimization of data networks," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 1535–1547, August 2006.

[43] K. Joseph and D. Raychaudhuri, "Throughput of unslotted direct-sequence spread-spectrum multiple-access channels with block FEC coding," *IEEE Trans. Inf. Theory*, vol. 41, pp. 1373–1378, September 1993.

[44] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall, 1993.

[45] S. M. Kay, *Fundamentals of Statistical Signal Processing - Detection Theory*. Prentice Hall, 1998.

[46] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek, "The Click modular router," *ACM Trans, Computer Systems*, vol. 18, pp. 263–297, August 2000.

[47] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 9, pp. 3037–3063, September 2005.

[48] J. Kuruvila, A. Nayak, and I. Stojmenovic, "Hop count optimal position based packet routing algorithms for ad hoc wireless networks with a realistic physical layer," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 1267–1275, June 2005.

[49] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, pp. 3062–3080, December 2004.

[50] J. N. Laneman, "Cooperative diversity in wireless networks: algorithms and architectures," *Ph. D. Thesis, Massachusets Institute of Technology*, September 2002.

[51] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2415–2425, October 2003.

[52] R. Laroia, S. Uppala, and J. Li, "Designing a mobile broadband wireless access network," *IEEE Signal Process. Mag.*, vol. 21, pp. 20–28, September 2004.

[53] P. Larsson., "Selection diversity forwarding in a multihop packet radio network with fading channel and capture," *ACM SIGMOBILE Mobile Computing and Commun. Review*, vol. 5, pp. 47–54, October 2001.

[54] R. Lin and A. P. Petropulu, "Cooperative transmission for random access wireless networks," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 3, pp. 373–376, Philadelphia, PA, March 19-23, 2005.

[55] R. Lin and A. P. Petropulu, "A new wireless network medium access protocol based on cooperation," in *Proc. Asilomar Conf. on Signals, Systems, Computers*, vol. 2, pp. 1922–1926, Pacific Grove, CA, November 7-10, 2004.

[56] Z. Liu, Y. Xin, and G. B. Giannakis, "Linear constellation precoding for OFDM with maximum multipath diversity and coding gains," *IEEE Trans. Commun.*, vol. 51, pp. 707–720, March 2003.

[57] S. H. Low, F. Paganini, and J. C. Doyle, "Internet congestion control," *IEEE Control Syst. Mag.*, vol. 22, pp. 28–43, February 2002.

[58] R. Loynes, "The stability of a queue with non-independent interarrival and service times," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, pp. 497–520, 1962.

[59] H. Lundgren, E. Nordstrom, and C. Tschudin, "The gray zone problem in IEEE 802.11b based ad hoc networks," *ACM SIGMOBILE Mobile Computing and Commun. Review*, vol. 3, pp. 104–105, July 2002.

[60] X. Ma and G. B. Giannakis, "Complex field coded MIMO systems: performance, rate, and trade-offs," *Wireless Commun. and Mobile Computing*, vol. 2, pp. 693–717, November 2002.

[61] D. A. Maltz, J. Broch, and D. B. Johnson, "Lessons from a full-scale multihop wireless ad hoc network testbed," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 8–15, February 2001.

[62] R. K. Morrow and J. S. Lehnert, "Packet throughput in slotted aloha DS/SSMA radio systems with random signature sequences," *IEEE Trans. Commun.*, vol. 40, pp. 1223–1230, July 1992.

[63] R. B. Myerson, *Game theory: analysis of conflict.* Harvard University Press, 1991.

[64] T. Nadeem and A. Agrawala, "IEEE 802.11 fragmentation-aware energy-efficient ad-hoc routing protocols," in *Proc. IEEE Int. Conf. Mobile Ad Hoc, Sensor Systems*, pp. 90–103, Fort Lauderdale, FL, October 25-27, 2004.

[65] NSF/ONR Workshop, "Cross-layer design in adaptive ad hoc networks: from signal processing to global networking,"

[66] U. of Minnesota, "The University of Minnesota: advancing the public good. Securing the universitys leadership position in the 21st century," *Report of the strategic positioning work group*, February 2005.

[67] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 1439–1451, August 2006.

[68] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications," *ACM SIGSAM*, pp. 37–44, March 1981.

[69] J. G. Proakis, *Digital Communications*. Mc Graw Hill, 4th ed., 2001.

[70] M. B. Pursley and D. J. Taipale, "Error probabilities for spread-spectrum packet radio with convolutional codes and viterbi decoding," *IEEE Trans. Commun.*, vol. 35, pp. 1–12, January 1987.

[71] R. Rao and A. Ephremides, "On the stability of interacting queues in a multi-access system," *IEEE Trans. Inf. Theory*, vol. 34, pp. 918–930, September 1988.

[72] T. S. Rappaport, *Wireless Communications*. Prentice Hall, 1996.

[73] A. Ribeiro, X. Cai, and G. B. Giannakis, "Opportunistic multipath for bandwidth-efficient cooperative networking," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 4, pp. 549–552, Montreal, Canada, May 17-21, 2004.

[74] A. Ribeiro, X. Cai, and G. B. Giannakis, "Symbol error probabilities for general cooperative links," in *Proc. IEEE Int. Conf. Commun.*, vol. 6, pp. 3369–3373, Paris, France, June 20-24, 2004.

[75] A. Ribeiro, X. Cai, and G. B. Giannakis, "Symbol error probabilities for general cooperative links," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1264–1273, May 2005.

[76] A. Ribeiro, X. Cai, and G. B. Giannakis, "Opportunistic multipath for bandwidth-efficient cooperative multiple access," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 2321–2327, September 2006.

[77] A. Ribeiro and G. B. Giannakis, "Fixed and Random Access Cooperative Networks," *EURASIP Newsletter*, vol. 17, pp. 3–24, March 2006.

[78] A. Ribeiro, G. B. Giannakis, and N. D. Sidiropoulos, "Stochastic routing in wireless multihop networks," *IEEE Signal Process. Mag.*, September 2006 (submitted). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[79] A. Ribeiro, Z.-Q. Luo, N. D. Sidiropoulos, and G. B. Giannakis, "A general optimization framework for stochastic routing in wireless multihop networks," *IEEE Trans. Signal Process.*, August 2006 (submitted). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[80] A. Ribeiro, Z.-Q. Luo, N. D. Sidiropoulos, and G. B. Giannakis, "A general optimization framework for stochastic routing in wireless multihop networks," in *Proc. Asilomar Conf. on Signals, Systems, Computers*, Pacific Grove, CA, October 29 - November 1, 2006 (to appear). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[81] A. Ribeiro, Z.-Q. Luo, N. D. Sidiropoulos, and G. B. Giannakis, "Modelling and optimization of stochastic routing for wireless multihop networks," in *Proc. IEEE INFOCOM*, Anchorage, Alaska, May 6-12, 2007 (submitted). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[82] A. Ribeiro, N. D. Sidiropoulos, and G. B. Giannakis, "Distributed routing algorithms for wireless multihop networks," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Honolulu, HI, April 15-20, 2006 (submitted). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[83] A. Ribeiro, N. D. Sidiropoulos, and G. B. Giannakis, "Optimal distributed stochastic routing algorithms for wireless multihop networks," *IEEE Trans. Commun.*, October 2006 (submitted). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[84] A. Ribeiro, N. D. Sidiropoulos, and G. B. Giannakis, "Achieving wireline random access throughput in wireless networking via user cooperation," in *Proc. IEEE Workshop on Signal Process. Advances in Wireless Commun.*, pp. 1033–1037, New York, NY, June 5-8, 2005.

[85] A. Ribeiro, N. D. Sidiropoulos, G. B. Giannakis, and Y. Yu, "Achieving wireline random access throughput in wireless networking via user cooperation," *IEEE Trans. Inf. Theory*, September 2006 (revised). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[86] A. Ribeiro, R. Wang, and G. B. Giannakis, "Linear complex-field coding for cooperative networking," in *Proc. of the first IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Process.*, pp. 48–51, Puerto Vallarta, Mexico, December 13-15, 2005.

[87] A. Ribeiro, R. Wang, and G. B. Giannakis, "Multi-source cooperation with full-diversity spectral-efficiency and controllable-complexity," in *Proc. IEEE Workshop on Signal Process. Advances in Wireless Commun.*, Cannes, France, July 2-5, 2006 (to appear). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[88] A. Ribeiro, R. Wang, and G. B. Giannakis, "Multi-source cooperation with full-diversity spectral-efficiency and controllable-complexity," *IEEE J. Sel. Areas Commun.*, March 2007 (to appear). Available at http:// www.ece.umn.edu/ users/ aribeiro/ research/ pubs.html.

[89] A. Ribeiro, Y. Yu, G. B. Giannakis, and N. D. Sidiropoulos, "Increasing the throughput of spread-aloha protocols via long PN spreading codes," in *Proc. IEEE Int. Conf. Commun.*, vol. 5, pp. 3628–3631, Seoul, Korea, May 16-20, 2005.

[90] E. M. Royer and C.-K. Toh, "A review of current routing protocols for ad-hoc mobile wireless networks," *IEEE Pers. Commun.*, vol. 6, pp. 46–55, April 1999.

[91] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic media access for multirate ad hoc networks," in *Proc. Int. ACM Conf. Mobile Computing, Networking*, pp. 24–35, Atlanta, GA, September 23-28, 2002.

[92] T. Sato, H. Okada, T. Yamazato, M. Katayama, and A. Ogawa, "Throughput analysis of DS/SSMA unslotted ALOHA system with fixed packet length," *IEEE J. Sel. Areas Commun.*, vol. 14, pp. 750–756, May 1996.

[93] A. Scaglione and Y. W. Hong, "Opportunistic large arrays: cooperative transmission in wireless multihop ad hoc networks to reach far distances," *IEEE Trans. Signal Process.*, vol. 51, pp. 2082–2092, August 2003.

[94] G. Scutari, S. Barbarossa, and D. Ludovici, "Cooperation diversity in multihop wireless networks using opportunistic driven multiple access," in *Proc. IEEE Workshop on Signal Process. Advances in Wireless Commun.*, Rome, Italy.

[95] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal multiplexing strategies for wideband meshed networks based on game theory part II: algorithms," *IEEE Trans. Signal Process.*, June 2006 (submitted).

[96] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part I: system description," *IEEE Trans. Commun.*, vol. 51, pp. 1927–1938, November 2003.

[97] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part II: implementation aspects and performance analysis," *IEEE Trans. Commun.*, vol. 51, pp. 1939–1948, November 2003.

[98] R. C. Shah, S. Wietholter, and A. Wolisz, "Modeling and analysis of opportunistic routing in low traffic scenarios," in *Proc. Symposium on Modeling and Optimization in Mobile, Ad Hoc, Wireless Netw.*, pp. 294–304, Trentino, Italy, April 3-7, 2005.

[99] R. C. Shah, S. Wietholter, A. Wolisz, and J. M. Rabaey, "When does opportunistic routing make sense?," in *Proc. IEEE Int. Conf. on Pervasive Computing and Commun.*, pp. 350–356, Kauai island, HI, March 8-12, 2005.

[100] O. Shalvi, "Multiple source cooperation diversity," *Zhi.-Quan Luo*, vol. 8, pp. 712–714, December 2004.

[101] M. K. Simon and M.-S. Alouini, *Digital Communications over Fading Channels.* Wiley-Interscience, 2000.

[102] E. S. Souza and J. A. Silvester, "Optimum transmission ranges in a direct sequence spread-spectrum multihop packet radio network," *IEEE J. Sel. Areas Commun.*, vol. 8, pp. 762–771, June 1990.

[103] I. Stojmenovic, A. Nayak, and J. Kuruvila, "Design guidelines for routing protocols in ad hoc and sensor networks with a realistic physical layer," *IEEE Commun. Mag.*, vol. 43, pp. 101–106, March 2005.

[104] J. F. Sturm, "Using Sedumi 1.02, a Matlab toolbox for optimization over symmetric cones," [Online]. Available at http://fewcal.kub.nl/sturm/software/sedumi.html.

[105] L. Tassiulas, "Adaptive back-pressure congestion control based on local information," *IEEE Trans. Autom. Control*, vol. 40, pp. 236–250, February 1995.

[106] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1067–1073, May 1997.

[107] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, pp. 1936–1948, December 1992.

[108] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.*, vol. 10, pp. 585–595, November 1999.

[109] M. K. Tsatsanis, R. Zhang, and S. Banerjee, "Network-assisted diversity for random access wireless networks," *IEEE Trans. Signal Process.*, vol. 48, pp. 702–711, March 2000.

[110] A. Tsirigos and Z. Haas, "Analysis of multipath routing. Part I: the effect on the packet delivery ratio," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 138–146, January 2004.

[111] B. Tsybakov and V. Mikhailov, "Ergodicity of a slotted aloha system," *ProbL Inf. Transm.*, vol. 15, April 1980. (translated from russian original in *Probl. Peredachi Inf.* 15, 4 (October-December 1979), 72-87).

[112] G. L. Turin, F. D. Clapp, T. L. Johnston, S. B. Fine, and D. Lavry, "A statistical model of urban multipath propagation," pp. 1–9, February 1972.

[113] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1277 – 1294, February 2002.

[114] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communication*. Addison-Wesley Wireless Comms. Series, 1995.

[115] R. Wang, W. Zhao, and G. B. Giannakis, "Multi-source cooperative networks with distributed convolutional coding," *IEEE Trans. Wireless Commun.*, September 2005 (submitted).

[116] R. Wang, W. Zhao, and G. B. Giannakis, "Distributed Trellis Coded Modulation for Multi-Source Cooperative Networks," in *Proc. of the IEEE Radio Wireless Symposium*, San Diego, CA, Jan. 17-19, 2006.

[117] R. Wang, W. Zhao, and G. B. Giannakis, "Multi-Source Cooperative Networks with Distributed Convolutional Coding," in *Proc. Asilomar Conf. on Signals, Systems, Computers*, Pacific Grove, CA, Oct. 30-Nov. 2, 2005.

[118] T. Wang, Y. Yao, and G. B. Giannakis, "Non-coherent distributed space-time processing for multiuser cooperative transmissions," *IEEE Transactions on Wireless Communications*, 2006 (to appear). See also *Proc. of Globecom Conf.*, St. Louis, MO, Nov. 28-Dec. 2, 2005.

[119] X. Wang and K. Kar, "Cross-layer rate optimization for proportional fairness in multihop wireless networks with random access," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 1548–1559, August 2006.

[120] X. Wang, Y. Yu, and A. Ribeiro, "Performance analysis of cooperative random access with long PN spreading codes," in *Proc. Asilomar Conf. on Signals, Systems, Computers*, pp. 499–503, Pacific Grove, CA, October 28 - November 1, 2005.

[121] Z. Wang, S. Zhou, and G. B. Giannakis, "Joint coding-precoding with low-complexity turbo-decoding," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 832–842, May 2004.

[122] Z. Wang and G. B. Giannakis, "A simple and general parameterization quantifying performance in fading channels," *IEEE Transactions on Communications*, vol. 51, pp. 1389–1398, Aug. 2003.

[123] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: where Fourier meets Shannon," *IEEE Signal Process. Mag.*, vol. 7, pp. 29–48, May 2000.

[124] Z. Wang and G. B. Giannakis, "Complex-field coding for OFDM over fading wireless channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 707–720, March 2003.

[125] Y. Xin, Z. Wang, and G. B. Giannakis, "Space-time diversity systems based on linear constellation precoding," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 294–309, March 2003.

[126] Y. Yao, X. Cai, and G. B. Giannakis, "On energy efficiency and optimum resource allocation in wireless relay transmissions," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2917–2927, November 2005.

[127] Y. Yu, A. Ribeiro, N. D. Sidiropoulos, and G. B. Giannakis, "Cooperative random access with long PN spreading codes," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 3, pp. 517–520, Philadelphia, PA, March 18-23, 2005.

[128] R. Zhang and M. S. Alouini, "Channel-aware inter-cluster routing protocol for wireless ad-hoc networks," in *Proc. Int. Symposium on Commun. Theory, Applications*, pp. 46–51, Ambleside, United Kingdom, July 2001.

[129] Q. Zhao and L. Tong, "Energy-efficient adaptive routing for ad hoc networks with time-varying heterogeneous traffic," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 5, pp. 801–804, Philadelphia, PA, March 18-23, 2005.

[130] W. Zhao and G. Giannakis, "Reduced complexity closest point decoding algorithms for random lattices," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 101–111, January.

[131] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1073–1096, May 2003.

[132] M. Zorzi and R. Rao, "Geographic Random Forwarding (GeRaF) for ad hoc and sensor networks: multihop performance," *IEEE Trans. Mobile Computing*, vol. 2, pp. 337–348, October - December 2003.