UNIVERSITY OF CALIFORNIA, SAN DIEGO

Topics in Network Communications

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Communication Theory and Systems)

by

Jillian Leigh Cannons

Committee in charge:

Professor Kenneth Zeger, Chair
Professor Massimo Franceschetti
Professor Laurence B. Milstein
Professor Lance Small
Professor Glenn Tesler
Professor Jack Wolf

2008

The dissertation of Jillian Leigh Cannons is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

____

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis is the culmination of research that would not have been possible without the efforts of many others. I would like to express my gratitude to Professor Ken Zeger who, through his approachability, patience, insight, and unbounded enthusiasm in our work, has allowed me to become a better scholar and has made this journey a success. Also, thank you to Professor Larry Milstein who so willingly lent his time, expertise, and humor to the final portion of this research. Thank you as well to Professors Massimo Franceschetti, Lance Small, Glenn Tesler, and Jack Wolf for serving on my committee and for their valued contributions. I am also indebted to my family, whose support and encouragement have never wavered throughout my many years of schooling. Finally, to my husband, Ryan Szypowski, who has walked this road alongside me, extending his arms when I faltered and ensuring that I did not lose sight of my dreams; his love, friendship, and belief in me have been absolute.

Parts of this thesis have been previously published as follows. The text of Chapter 2, in full, is a reprint of the material as it appears in Jillian Cannons, Randall Dougherty, Chris Freiling, and Kenneth Zeger, "Network Routing Capacity," *IEEE Transactions on Information Theory*, vol. 52, no. 3, March 2006, with copyright held by the IEEE.[1] The text of Chapter 3, in full, is a reprint of the material as it appears in Jillian Cannons and Kenneth Zeger, "Network Coding Capacity With a Constrained Number of Coding Nodes," *IEEE Transactions on Information Theory*, vol. 54, no. 3, March 2008, with copyright held by the IEEE.[1] With the exception of the appendix, the text of Chapter 4, in full, has been submitted for publication as Jillian Cannons, Laurence B. Milstein, and Kenneth Zeger, "An Algorithm for Wireless Relay Placement," *IEEE Transactions on Wireless Communications*, submitted August 4, 2008. In all three cases I was a primary researcher and the co-author Kenneth Zeger directed and supervised the research which forms the basis of this dissertation. Co-authors Randall Dougherty and Chris Freiling contributed to the research on network routing capacity, while co-author Laurence B. Milstein contributed to the research on wireless relay placement.

VITA

2000	B.Sc. in Computer Engineering *with Distinction* University of Manitoba
2000-2002	Teaching Assistant University of Illinois at Urbana-Champaign
2002	M.S. in Electrical Engineering University of Illinois at Urbana-Champaign
2002-2003, 2006	Teaching Assistant University of California, San Diego
2003-2008	Graduate Student Researcher University of California, San Diego
2008	Ph.D. in Electrical Engineering (Communication Theory and Systems) University of California, San Diego

PUBLICATIONS

J. Cannons, L. B. Milstein, and K. Zeger, "An Algorithm for Wireless Relay Placement," *IEEE Transactions on Wireless Communications*, submitted August 4, 2008.

J. Cannons and K. Zeger, "Network Coding Capacity With a Constrained Number of Coding Nodes," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1287-1291, March 2008.

J. Cannons and K. Zeger, "Network Coding Capacity with a Constrained Number of Coding Nodes." *Proceedings of the 44th Annual Allerton Conference on Communications, Control, and Computing*, 3 pages, Allerton Park, IL, September 27-29, 2006 (Invited).

J. Cannons, R. Dougherty, C. Freiling, and K. Zeger, "Network Routing Capacity," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 777-788, March 2006.

J. Cannons, R. Dougherty, C. Freiling, and K. Zeger, "Network Routing Capacity," *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pp. 11-13, Adelaide, Australia, September 4-9, 2005.

J. Cannons and P. Moulin, "Design and Statistical Analysis of a Hash-Aided Image Watermarking System," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1393-1408, October 2004.

J. Cannons and W. Kinsner "Modelling of Lightning Discharge Patterns as Observed from Space," *Proceedings of the 12th International Conference on Mathematical and Computer Modeling and Scientific Computing*, Chicago, Illinois, August 2-4, 1999.

ABSTRACT OF THE DISSERTATION

Topics in Network Communications

by

Jillian Leigh Cannons
Doctor of Philosophy in Electrical Engineering
(Communication Theory and Systems)

University of California San Diego, 2008

Professor Kenneth Zeger, Chair

This thesis considers three problems arising in the study of network communications. The first two relate to the use of network coding, while the third deals with wireless sensor networks.

In a traditional communications network, messages are treated as physical commodities and are routed from sources to destinations. Network coding is a technique that views data as information, and thereby permits coding between messages. Network coding has been shown to improve performance in some networks. The first topic considered in this thesis is the routing capacity of a network. We formally define the routing and coding capacities of a network, and determine the routing capacity for various examples. Then, we prove that the routing capacity of every network is achievable and rational, we present an algorithm for its computation, and we prove that every rational number in $(0, 1]$ is the routing capacity of some solvable network. We also show that the coding capacity of a network is independent of the alphabet used.

The second topic considered is the network coding capacity under a constraint on the total number of nodes that can perform coding. We prove that every non-negative,

monotonically non-decreasing, eventually constant, rational-valued function on the non-negative integers is equal to the capacity as a function of the number of allowable coding nodes of some direct acyclic network.

The final topic considered is the placement of relays in wireless sensor networks. Wireless sensor networks typically consist of a large number of small, power-limited sensors which collect and transmit information to a receiver. A small number of relays with additional processing and communications capabilities can be strategically placed to improve system performance. We present an algorithm for placing relays which attempts to minimize the probability of error at the receiver. We model communication channels with Rayleigh fading, path loss, and additive white Gaussian noise, and include diversity combining at the receiver. For certain cases, we give geometric descriptions of regions of sensors which are optimally assigned to the same, fixed relays. Finally, we give numerical results showing the output and performance of the algorithm.

Chapter 1

Introduction

The study of data communications was revolutionized in 1948 by Shannon's seminal paper "A Mathematical Theory of Communication" [26]. Shannon's work introduced the framework of information theory (e.g., see [8]), and established both the rate at which data can be compressed and the rate at which data can be transmitted over a noisy channel. Equipped with this knowledge, the field of digital communications (e.g., see [24]) addresses the question of how data should be transmitted. The study of network communications builds further upon these foundations by examining information exchange amongst members of a set of sources and receivers.

This thesis considers three topics in two subfields of network communications. The first two relate to the use of network coding (e.g, see [31]), which is a technique that permits coding between streams of transmitted information. The third topic deals with wireless sensor networks (e.g, see [17]), which typically are groups of small, data-collecting nodes that transmit information to a receiver. Both of these areas of network communications have emerged in the last decade and have since garnered considerable attention.

1.1 Network Coding

A communications network can be modeled by a directed, acyclic multigraph. A subset of the nodes in the graph are source nodes, which emit source node messages. Similarly, a subset of the nodes are sink nodes, which demand specific source node messages. Each source message is taken to be a vector of k symbols, while each edge can carry a vector of n symbols. Traditionally, network messages are treated as physical commodities, which are routed throughout the network without replication or alteration. Conversely, the field of network coding views network messages as information, which can be copied and transformed by any node within the network. Specifically, the value on each outgoing edge of a node is some function of the values on its incoming edges (and emitted messages if it is a source). A goal in network coding is to determine a coding function for each edge in the network such that each sink can perform decoding operations to determine its desired source messages. Ahlswede, Cai, Li, and Yeung [1] demonstrated that there exist networks for which network coding (as opposed to simply routing) is required to satisfy the sink demands. Figure 1.1 gives two copies of a network where source node 1 emits message $\mathbf{x}$, source node 2 emits message $\mathbf{y}$, sink node 5 demands messages $\mathbf{x}$ and $\mathbf{y}$, and sink node 6 demands messages $\mathbf{x}$ and $\mathbf{y}$. The left version depicts an attempt to provide a routing solution, however the bottleneck between nodes 3 and 4 prohibits both messages $\mathbf{x}$ and $\mathbf{y}$ from arriving at both sinks. (In the given attempt, the demands of sink 5 are not met.) The right version demonstrates a solution using network coding, where the edge between nodes 3 and 4 carries the sum of messages $\mathbf{x}$ and $\mathbf{y}$. Both sinks can decode both messages using subtraction. This solution is valid for messages drawn from any group with group operator "$+$". Figure 1.2 gives a numerical example of the same network coding solution with message components from the binary field $\mathbb{Z}_2$ with "$+$" being addition modulo 2 (i.e., the XOR function). In the depicted example, both the messages and the edges are of vector dimension $k = n = 2$.

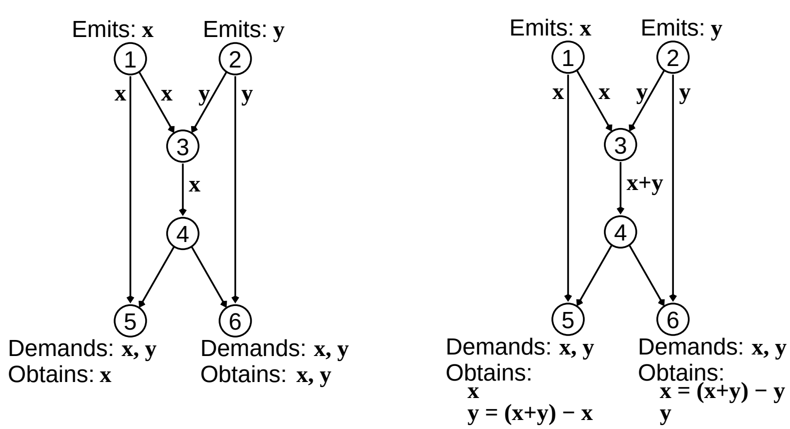

Figure 1.1: Example network with source nodes 1 and 2 and sink nodes 5 and 6. Left: Only routing is permitted. Right: Network coding is permitted.

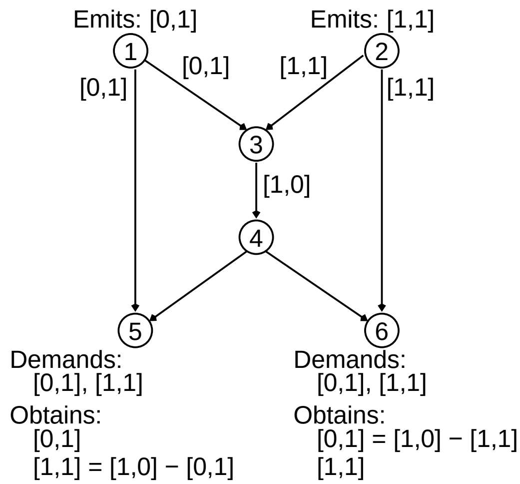

Figure 1.2: Numerical example of the network coding solution in Figure 1.1.

We define the coding capacity of a network to be the largest ratio of source message vector dimension to edge vector dimension for which there exist edge functions allowing sink demands to be satisfied. Analogously, we define the routing capacity for the case when network nodes are only permitted to perform routing, and the linear coding capacity for the case when only linear edge functions are permitted. Larger capacity values correspond to better performance, and comparing the routing capacity to the coding capacity illustrates the benefit of network coding over routing. It is known that the linear coding capacity can depend on the alphabet size [9], whereas the routing capacity is trivially independent of the alphabet. We prove in Chapter 2 that the general coding capacity is independent of the alphabet used. It is not presently known whether the coding capacity or the linear coding capacity must be rational numbers, nor if the linear coding capacity is always achievable. It has recently been shown, however, that the (general) coding capacity of a network need not be achievable [10]. We prove in Chapter 2 that the routing capacity of every network is achievable (and therefore is also rational). The computability of coding capacities is in general an unsolved problem. For example, it is presently not known whether there exists an algorithm for determining the capacity or the linear coding capacity of a network. We prove in Chapter 2 that the routing capacity of a network is computable, by explicitly demonstrating a linear program solution. Chapter 2 is reprint of paper appearing in the IEEE Transactions on Information Theory.

It is also interesting to consider the number of coding nodes required to achieve the coding capacity of a network. A similar problem is to determine the number of coding nodes needed to satisfy the sink demands for the case when messages are of the same vector dimension as edges. The number of required coding nodes in both problems can in general range anywhere from zero up to the total number of nodes in the network. The later problem has been examined previously by Langberg, Sprintson, and Bruck [19], Tavory, Feder, and Ron [12], Fragouli and Soljanin [13], Bhattad, Ratnakar, Koetter, and Narayanan [3], and Wu, Jain, and Kung [30] for the special case of networks containing only a single source and with all sinks demanding all source messages. We study the related (and more general) problem of how the coding capacity varies as a function of the number of allowable coding nodes. For example, the network in Figure 1.1 has capacity $1/2$ when no coding nodes are permitted (achievable by taking message dimension 1 and edge dimension 2) and capacity 1 when one or more coding nodes are permitted. In Chapter 3 we show that nearly any non-decreasing function is the capacity as a function of the number of allowable coding nodes of some network. Thus, over all directed, acyclic networks, arbitrarily large amounts of coding gain can be attained by using arbitrarily-sized node subsets for coding. Chapter 3 is reprint of paper appearing in the IEEE Transactions on Information Theory.

1.2 Wireless Sensor Networks

A wireless sensor network is a possibly large group of small, power-limited sensors distributed over a geographic area. The sensors collect information which is transmitted to a receiver for further analysis. Applications of such networks include the monitoring of environmental conditions, the tracking of moving objects, and the detection of events of interest. A small number of radio relays with additional processing and communications capabilities can be strategically placed in a wireless sensor network to improve system performance. A sample wireless sensor network is shown in Figure 1.3 where the sensors are denoted by circles, the relays by triangles, and the receiver by a square. Two important problems are to position the relays and to determine, for each sensor, which relay should rebroadcast its signal.

In order to compare various relay placements and sensor assignments, a communications model and an optimization goal must be determined. We assume transmission

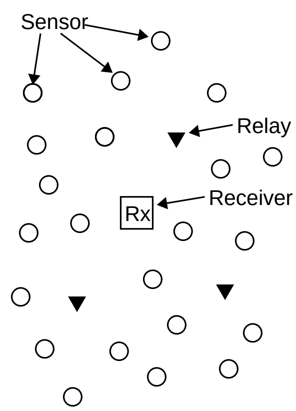

Figure 1.3: A wireless sensor network with sensors denoted by circles, relays by triangles, and the receiver by a square.

occur using binary phase shift keying (BPSK) in which a single bit is sent by modulating a pulse with a cosine wave. The magnitude of the transmitted signal diminishes with the distance traveled, which is known as path loss. Furthermore, since transmissions occur wirelessly, a given transmitted signal may traverse multiple paths to the destination (e.g., direct transmission versus bouncing off a building wall), causing the receiver to obtain multiple copies of the signal. This effect is known as multi-path fading and is modeled using a random variable. Finally, additive white Gaussian noise (AWGN) is also present at receiving antennae. We consider relays using either the amplify-and-forward or the decode-and-forward protocol. An amplify-and-forward relay generates an outgoing signal by multiplying an incoming signal by a gain factor. A decode-and-forward relay generates an outgoing signal by making a hard decision on the value of the bit represented by an incoming signal, and transmits a regenerated signal using the result. Each sensor in the network transmits information to the receiver both directly and through a relay path. The receiver combines the two received signals to achieve transmission diversity. We assume transmissions are performed using a slotted mechanism such as time division multiple access (TDMA) so that there is ideally no transmission interference. Figure 1.4 shows the example wireless sensor network over a sequence of time slots with transmission occurring using TDMA and single-hop relay paths. Using this network model, we attempt to position the relays and assign sensors to them in order to minimize the average probability of error at the receiver.

Previous studies of relay placement have considered various optimization criteria

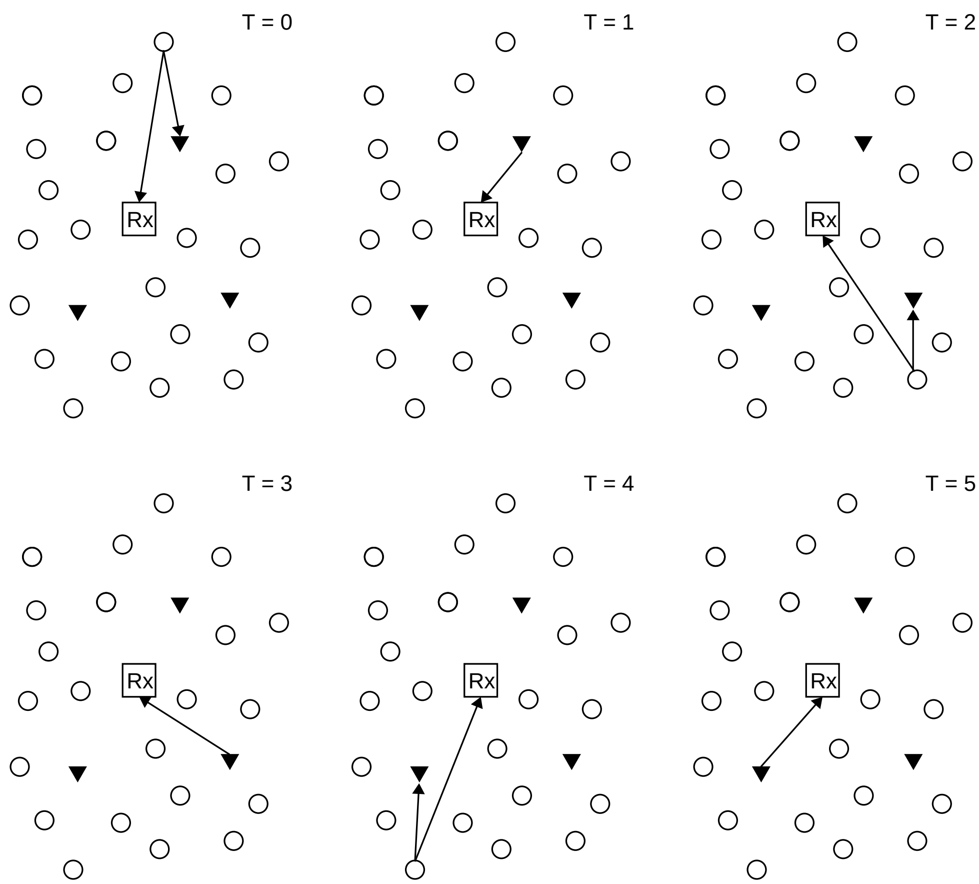

Figure 1.4: Transmissions in a wireless sensor network over six time slots.

and communication models. For example, coverage, lifetime, energy usage, error probability, outage probability, or throughput were focused on by Balam and Gibson [2]; Chen and Laneman [4]; Chen, Wang, and Liang [5]; Cho and Yang [6]; Cortés, Martiínez, Karataş, and Bullo [7]; Ergen and Varaiya [11]; Hou, Shi, Sherali, and Midkiff [15]; Iranli, Maleki, and Pedram [16]; Koutsopoulos, Toumpis, and Tassiulas [18]; Liu and Mohapatra [20]; Ong and Motani [22]; Mao and Wu [21]; Suomela [28]; Tan, Lozano, Xi, and Sheng [29]; Pan, Cai, Hou, Shi, and Shen [23]; Sadek, Han, and Liu [25]; So and Liang [27]. The communications and/or network models used are typically simplified by techniques such as assuming error-free communications, assuming transmission energy is an increasing function of distance, assuming single sensor networks, assuming single relay networks, and excluding diversity.

In Chapter 4 we present an algorithm that determines relay placement and assigns each sensor to a relay. The algorithm has some similarity to a source coding design technique known as the Lloyd algorithm (e.g., see [14]). We describe geometrically, with respect to fixed relay positions, the sets of locations in the plane in which sensors are (optimally) assigned to the same relay, and give performance results based on these analyses and using numerical computations. Chapter 4 has been submitted as a paper to the IEEE Transactions on Wireless Communications.

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204 – 1216, 2000.

[2] J. Balam and J. D. Gibson, "Adaptive event coverage using high power mobiles over a sensor field," in *Proceedings of IEEE 60th Vehicular Technology Conference*, Los Angeles, CA, September 26 - 29, 2004, pp. 4611 – 4615.

[3] K. Bhattad, N. Ratnakar, R. Koetter, and K. R. Narayanan, "Minimal network coding for multicast," in *Proceedings of the 2005 IEEE International Symposium on Information Theory*, Adelaide, Australia, September 4-9, 2005.

[4] D. Chen and J. N. Laneman, "Modulation and demodulation for cooperative diversity in wireless systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1785 – 1794, 2006.

[5] Y. Chen, Z. Wang, and J. Liang, "Automatic dynamic flocking in mobile actuator sensor networks by central Voronoi tessellations," in *Proceedings of the IEEE International Conference on Mechatronics and Automation*, Niagara Falls, Canada, July 29 - August 1, 2005, pp. 1630 – 1635.

[6] W. Cho and L. Yang, "Energy and location optimization for relay networks with differential modulation," in *Proceedings of the 25th Army Science Conference*, Orlando, FL, November 27-30, 2006.

[7] J. Cortés, S. Martínez, T. Karataş, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243 – 255, 2004.

[8] T. Cover and J. Thomas, *Elements of Information Theory*. U.S.A.: John Wiley & Sons, Inc., 1991.

[9] R. Dougherty, C. Freiling, and K. Zeger, "Linearity and solvability in multicast networks," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2243 – 2256, 2004.

[10] ——, “Unachievability of network coding capacity,” *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2365 – 2372, 2006.

[11] S. C. Ergen and P. Varaiya, “Optimal placement of relay nodes for energy efficiency in sensor networks,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, Istanbl, Turkey, June 11-15, 2006, pp. 3473 – 3479.

[12] M. Feder, D. Ron, and A. Tavory, “Bounds on linear codes for network multicast,” *Electronic Colloquium on Computational Complexity (ECCC)*, pp. 1 – 9, 2003, report 33.

[13] C. Fragouli and E. Soljanin, “Information flow decomposition for network coding,” *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 829 – 848, 2006.

[14] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic Publishers, 1992.

[15] Y. T. Hou, Y. Shi, H. D. Sherali, and S. F. Midkiff, “On energy provisioning and relay node placement for wireless sensor networks,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 5, pp. 2579 – 2590, 2005.

[16] A. Iranli, M. Maleki, and M. Pedram, “Energy efficient strategies for deployment of a two-level wireless sensor network,” in *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, San Diego, CA, August 8-10, 2005, pp. 233 – 238.

[17] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*. Hoboken, NJ: John Wiley & Sons, 2005.

[18] I. Koutsopoulos, S. Toumpis, and L. Tassiulas, “On the relation between source and channel coding and sensor network deployment,” in *Proceedings of the International Workshop on Wireless Ad-hoc Networks (IWWAN)*, London, England, May 23-26, 2005.

[19] M. Langberg, A. Sprintson, and J. Bruck, “The encoding complexity of network coding,” *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2386 – 2397, 2006.

[20] X. Liu and P. Mohapatra, “On the deployment of wireless sensor nodes,” in *Proceedings of the 3rd International Workshop on Measurement, Modeling, and Performance Analysis of Wireless Sensor Networks (SenMetrics)*, San Diego, CA, July 21, 2005, pp. 78 – 85.

[21] Y. Mao and M. Wi, “Coordinated sensor deployment for improving secure communication and sensing coverage,” in *Proceedings of the Third ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN)*, Alexandria, VA, November 7, 2005.

[22] L. Ong and M. Motani, "On the capacity of the single source multiple relay single destination mesh network," *Ad Hoc Networks*, vol. 5, no. 6, pp. 786 – 800, 2007.

[23] J. Pan, L. Cai, Y. T. Hou, Y. Shi, and S. X. Shen, "Optimal base-station locations in two-tiered wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 4, no. 5, pp. 458 – 473, 2005.

[24] J. G. Proakis, *Digital Communications*, 4th ed. New York, NY: McGraw-Hill, 2001.

[25] A. K. Sadek, Z. Han, and K. J. Liu, "An efficient cooperation protocol to extend coverage area in cellular networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, Las Vegas, NV, April 3-6, 2006, pp. 1687 – 1692.

[26] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379 – 423 and 623 – 656, 1948.

[27] A. So and B. Liang, "Exploiting spatial diversity in rate adaptive WLANs with relay infrastructure," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*, St. Louis, MO, November 28 - December 2, 2005.

[28] J. Suomela, "Approximating relay placement in sensor networks," in *Proceedings of the 3rd ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor and Ubiquitous Networks (PE-WASUN)*, Terromolinos, Spain, October 2-6, 2006, pp. 145 – 148.

[29] J. Tan, O. M. Lozano, N. Xi, and W. Sheng, "Multiple vehicle systems for sensor network area coverage," in *Proceedings of the 5th World Conference on Intelligent Control and Automation*, Hangzhou, P. R. China, June 15-19, 2004, pp. 4666 – 4670.

[30] Y. Wu, K. Jain, and S.-Y. Kung, "A unification of network coding and tree packing (routing) theorems," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2398 – 2409, 2006.

[31] R. W. Yeung, *A First Course in Information Theory*. New York, NY: Kluwer Academic, 2002.

Chapter 2

Network Routing Capacity

Abstract

We define the routing capacity of a network to be the supremum of all possible fractional message throughputs achievable by routing. We prove that the routing capacity of every network is achievable and rational, we present an algorithm for its computation, and we prove that every rational number in $(0, 1]$ is the routing capacity of some solvable network. We also determine the routing capacity for various example networks. Finally, we discuss the extension of routing capacity to fractional coding solutions and show that the coding capacity of a network is independent of the alphabet used.

2.1 Introduction

A communications network is a finite, directed, acyclic multigraph over which messages can be transmitted from source nodes to sink nodes. The messages are drawn from a specified alphabet, and the edges over which they are transmitted are taken to be error-free, cost-free, and of zero-delay. Traditionally, network messages are treated as physical commodities, which are routed throughout the network without replication or alteration. However, the emerging field of network coding views the messages as information, which can be copied and transformed by any node within the network. Network coding permits each outgoing edge from a node to carry some function of the data received on the incoming edges of the node. A goal in using network coding is to determine a set of edge functions that allow all of the sink node demands to be satisfied. If such a set of functions exists, then the network is said to be *solvable*, and the functions are called a *solution*. Otherwise the network is said to be *unsolvable*.

A solution to a network is said to be a *routing solution* if the output of every edge function equals a particular one of its inputs. A solution to a network is said to be a *linear solution* if the output of every edge function is a linear combination of its inputs, where linearity is defined with respect to some underlying algebraic structure on the alphabet, usually a finite field or ring. Clearly, a routing solution is also a linear solution.

Network messages are fundamentally scalar quantities, but it is also useful to consider blocks of multiple scalar messages from a common alphabet as message vectors. Such vectors may correspond to multiple time units in a network. Likewise, the data transmitted on each network edge can also be considered as vectors. *Fractional coding* refers to the general case where message vectors differ in dimension from edge data vectors (e.g., see [2]). The coding functions performed at nodes take vectors as input on each in-edge and produce vectors as output on each out-edge. A *vector linear solution* has edge functions which are linear combinations of vectors carried on in-edges to a node, where the linear combination coefficients are matrices over the same alphabet as the input vector components. In a *vector routing solution* each edge function copies a collection of components from input edges into a single output edge vector.

For any set of vector functions which satisfies the demands of the sinks, there is a corresponding scalar solution (by using a Cartesian product alphabet). However, it is known that if a network has a vector routing solution, then it does not necessarily have a scalar routing solution. Similarly, if a network has a vector linear solution, then it does not necessarily have a scalar linear solution [16].

Ahlswede, Cai, Li, and Yeung [1] demonstrated that there exist networks with (linear) coding solutions but with no routing solutions, and they gave necessary conditions for solvability of multicast networks (networks with one source and all messages demanded by all sink nodes).

Li, Yeung, and Cai [15] proved that any solvable multicast network has a scalar linear solution over some sufficiently large finite field alphabet.

For multicast networks, it is known that solvability over a particular alphabet does not necessarily imply scalar linear solvability over the same alphabet (see examples in [4], [18], [16], [20]). For non-multicast networks, it has recently been shown that solvability does not necessarily imply vector linear solvability [5].

Rasala Lehman and Lehman [19] have noted that for some networks, the size of the alphabet needed for a solution can be significantly reduced if the solution does not operate at the full capacity of the network. In particular, they demonstrated that, for certain networks, fractional coding can achieve a solution where the ratio of edge capacity n to message vector dimension k is an arbitrarily small amount above one. The observations in [19] suggest many important questions regarding network solvability using fractional coding.

In the present paper, we focus on such fractional coding for networks in the special case of routing[1]. We refer to such coding as *fractional routing*. Specifically, we consider message vectors whose dimension may differ from the dimension of the vectors carried on edges. Only routing is considered, so that at any node, any set of components of the node's input vectors may be sent on the out-edges, provided the edges' capacities are not exceeded.

We define a quantity called the *routing capacity* of a network, which characterizes the highest possible capacity obtainable from a fractional routing solution to a network[2]. The routing capacity is the the supremum of ratios of message dimension to edge capacity for which a routing solution exists. Analogous definitions can be made of the (general) coding capacity over all (linear and non-linear) network codes and the linear coding capacity over all linear network codes. These definitions are with respect to the specified alphabet and are for general networks (e.g., they are not restricted to multicast networks).

---

[1]Whereas the present paper studies networks with directed edges, some results on fractional coding were obtained by Li et al. [13], [14] for networks with undirected (i.e., bidirectional) edges.

[2]Determining the routing capacity of a (directed) network relates to the maximum throughput problem in an undirected network in which multiple multicast sessions exist (see Li et al. [13], [14]), with each demanded message being represented by a multicast group. In the case where only a single multicast session is present in the network, determining the routing capacity corresponds to fractional directed Steiner tree packing, as considered by Wu, Chou, and Jain [23] and, in the undirected case, by Li et al. [13], [14]. In the case where the (directed) network has disjoint demands (i.e., when each message is only demanded by a single sink), determining the routing capacity resembles the maximum concurrent multicommodity flow problem [22].

It is known that the linear coding capacity (with respect to a finite field alphabet) can depend on the alphabet size [5] whereas the routing capacity is trivially independent of the alphabet. We prove here, however, that the general coding capacity is independent of the alphabet used.

It is not presently known whether the coding capacity or the linear coding capacity of a network must be rational numbers. Also, it is not presently known if the linear coding capacity of a network is always achievable. It has recently been shown, however, that the (general) coding capacity of a network need not be achievable [6]. We prove here that the routing capacity of every network is achievable (and therefore is also rational). We also show that every rational number in $(0, 1]$ is the routing capacity of some solvable network.

The computability of coding capacities is in general an unsolved problem. For example, it is presently not known whether there exists an algorithm for determining the coding capacity or the linear coding capacity (with respect to a given alphabet size) of a network. We prove here that the routing capacity is indeed computable, by explicitly demonstrating a linear program solution. We do not attempt to give a low complexity or efficient algorithm, as our intent is only to establish the computability of routing capacity.

Section 2.2 gives formal definitions of the routing capacity and related network concepts. Section 2.3 determines the routing capacity of a variety of sample networks in a semi-tutorial fashion. Section 2.4 proves various properties of the routing capacity, including the result that the routing capacity is achievable and rational. Section 2.5 gives the construction of a network with a specified routing capacity. Finally, Section 2.6 defines the coding capacity of a network and shows that it is independent of the alphabet used.

2.2 Definitions

A *network* is a finite, directed, acyclic multigraph, together with non-empty sets of source nodes, sink[3] nodes, source node messages, and sink node demands. Each message is an arbitrary element of a fixed finite alphabet and is associated with exactly one source node, and each demand at a sink node is a specification of a specific source message that needs to be obtainable at the sink. A network is *degenerate* if there exists a source message

[3] Although the terminology "sink" in graph theory indicated a node with no out-edges, we do not make that restriction here. We merely refer to a node which demands at least one message as a sink.

demanded at a particular sink, but with no directed path through the graph from the source to the sink.

Each edge in a network carries a vector of symbols from some alphabet. The maximum allowable dimension of these vectors is called the *edge capacity*. (If an edge carries no alphabet symbols, it is viewed as carrying a vector of dimension zero.) Note that a network with nonuniform, rational-valued edge capacities can always be equivalently modeled as a network with uniform edge capacities by introducing parallel edges. For a given finite alphabet, an *edge function* is a mapping, associated with a particular edge (u, v), which takes as inputs the edge vector carried on each in-edge to the node u and the source messages generated at node u, and produces an output vector to be carried on the edge (u, v). A *decoding function* is a mapping, associated with a message demanded at a sink, which takes as inputs the edge vector carried on each in-edge to the sink and the source messages generated at the sink, and produces an output vector hopefully equal to the demanded message.

A *solution* to a network for a given alphabet is an assignment of edge functions to a subset of edges and an assignment of decoding functions to all sinks in the network, such that each sink node obtains all of its demands. A network is *solvable* if it has a solution for some alphabet. A network solution is a *vector routing solution* if every edge function is defined so that each component of its output is copied from a (fixed) component of one of its inputs. (So, in particular, no "source coding" can occur when generating the outputs of source nodes.) It is clear that vector routing solutions do not depend on the chosen alphabet. A solution is *reducible* if it has at least one edge function which, when removed, still yields a solution. A vector solution is *reducible* if it has at least one component of at least one edge function which, when removed, still yields a vector solution.

A (k, n) *fractional routing solution* of a network is a vector routing solution that uses messages with k components and edges with capacity n, with $k, n \geq 1$. Note that if a network is solvable then it must have a (coding) solution with $k = n = 1$. A (k, n) fractional routing solution is *minimal* if it is not reducible and if no (k, n') fractional routing solution exists for any $n' < n$. Solvable networks may or may not have routing solutions. However, every nondegenerate network has a (k, n) fractional routing solution for some k and n. In fact, it is easy to construct such a solution by choosing $k = 1$ and n equal to the total number of messages in the network, since then every edge has enough capacity to carry every message that can reach it from the sources.

The ratio k/n in a (k, n) fractional routing solution quantifies the capacity of the solution and the rational number k/n is said to be an *achievable routing rate* of the network. Define the set

$$U = \{r \in \mathbb{Q} : r \text{ is an achievable routing rate}\}.$$

The *routing capacity* of a network is the quantity

$$\epsilon = \sup U.$$

If a network has no achievable routing rate then we make the convention that $\epsilon = 0$. It is clear that $\epsilon = 0$ if and only if the network is degenerate. Also, $\epsilon < \infty$ (e.g., since k/n is trivially upper bounded by the number of edges in the network). Note that the supremum in the definition of ϵ can be restricted to achievable routing rates associated with minimal routing solutions. The routing capacity is said to be *achievable* if it is an achievable routing rate. Note that an achievable routing capacity must be rational. A fractional routing solution is said to *achieve* the routing capacity if the routing rate of the solution is equal to the routing capacity.

Intuitively, for a given network edge capacity, the routing capacity bounds the largest message dimension for which a routing solution exists. If $\epsilon = 0$, then at least one sink has an unsatisfied demand, which implies that no path between the sink and the source emitting the desired message exists. If $\epsilon \in (0, 1)$, then the edge capacities need to be inflated with respect to the message dimension to satisfy the demands of the sinks. If $\epsilon = 1$, then it will follow from results in this paper that a fractional routing solution exists where the message dimensions and edge capacities are identical. If $\epsilon > 1$, then the edge capacities need not even be as large as the message dimension to satisfy the demands of the sinks. Finally, if a network has a routing solution, then the routing capacity of the network satisfies $\epsilon \geq 1$.

2.3 Routing Capacity of Example Networks

To illustrate the concept of the routing capacity, a number of examples are now considered. For each example in this section, let k be the dimension of the messages and let n be the capacity of the edges. All figures in this section have graph nodes labeled by positive integers. Any node labeled by integer i is referred to as n_i. Also, any edge connecting nodes i and j is referred to as $e_{i,j}$ (instead of the usual notation (i,j)), as is the message vector carried by the edge. The distinction between the two meanings of $e_{i,j}$ is made clear in each such instance.

Example 2.3.1. (See Figure 2.1.)

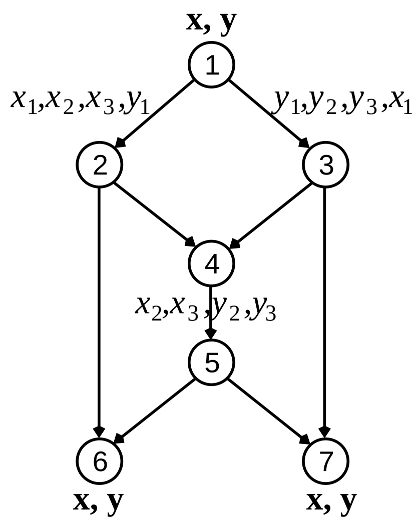

Figure 2.1: The multicast network $\mathcal{N}_1$ whose routing capacity is $3/4$.

The single source produces two messages which are both demanded by the two sinks. The network has no routing solution but does have a linear coding solution [1]. The routing capacity of this multicast network is $\epsilon = 3/4$.

Proof. In order to meet the sink node demands, each of the $2k$ message components must be carried on at least two of the three edges $e_{1,2}, e_{1,3}$, and $e_{4,5}$ (because deleting any two of these three edges would make at least one of the sinks unreachable from the source). Hence, we have the requirement $2(2k) \leq 3n$, for arbitrary k and n. Hence $\epsilon \leq 3/4$.

Now, let $k = 3$ and $n = 4$, and route the messages as follows:

$$e_{1,2} = e_{2,6} = (x_1, x_2, x_3, y_1)$$
$$e_{1,3} = e_{3,7} = (y_1, y_2, y_3, x_1)$$

$$e_{2,4} = (x_2, x_3)$$
$$e_{3,4} = (y_2, y_3)$$
$$e_{4,5} = (x_2, x_3, y_2, y_3)$$
$$e_{5,6} = (y_2, y_3)$$
$$e_{5,7} = (x_2, x_3).$$

This is a fractional routing solution to $\mathcal{N}_1$. Thus, $3/4$ is an achievable routing rate of $\mathcal{N}_1$, so $\epsilon \geq 3/4$. ■

Example 2.3.2. (See Figure 2.2.)

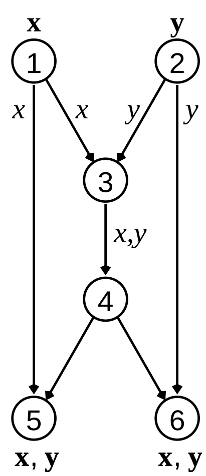

Figure 2.2: The network $\mathcal{N}_2$ whose routing capacity is $1/2$.

Each of the two sources emits a message and both messages are demanded by the two sinks. The network has no routing solution but does have a linear coding solution (similar to Example 2.3.1). The routing capacity of this network is $\epsilon = 1/2$.

Proof. The only path over which message $\mathbf{x}$ can be transmitted from source n_1 to sink n_6 is n_1, n_3, n_4, n_6. Similarly, the only path feasible for the transmission of message $\mathbf{y}$ from source n_2 to sink n_5 is n_2, n_3, n_4, n_5. Thus, there must be sufficient capacity along edge $e_{3,4}$ to accommodate both messages. Hence, we have the requirement $2k \leq n$, yielding $k/n \leq 1/2$ for arbitrary k and n. Thus, $\epsilon \leq 1/2$.

Now, let $k = 1$ and $n = 2$, and route the messages as follows:

$$e_{1,5} = e_{1,3} = e_{4,6} = (\mathbf{x})$$

$$e_{2,6} = e_{2,3} = e_{4,5} = (\mathbf{y})$$
$$e_{3,4} = (\mathbf{x}, \mathbf{y}).$$

This is a fractional routing solution to $\mathcal{N}_2$. Thus, $1/2$ is an achievable routing rate of $\mathcal{N}_2$, so $\epsilon \geq 1/2$. ■

Example 2.3.3. (See Figure 2.3.)

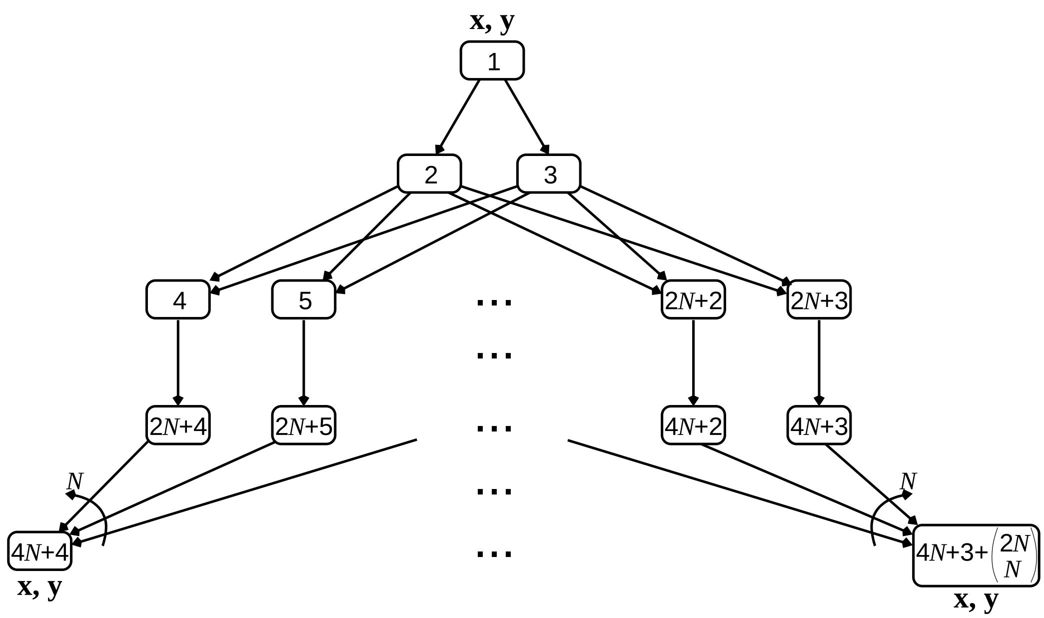

Figure 2.3: The multicast network $\mathcal{N}_3$ whose routing capacity is $N/(N+1)$.

The network $\mathcal{N}_3$ contains a single source n_1 with two messages, $\mathbf{x}$ and $\mathbf{y}$. The second layer consists of two nodes, n_2 and n_3. The third and fourth layers each contain $2N$ nodes. The bottom layer contains $\binom{2N}{N}$ sink nodes, where each such node is connected to a distinct set of N nodes from the fourth layer. Each of these sink nodes demands both source messages. The network has no routing solution but does have a linear coding solution for $N \geq 2$ (since the network is multicast and the minimum cut size is 2 for each sink node [15]). The routing capacity of this network is $\epsilon = N/(N+1)$.

Proof. Let $\mathbf{D}$ be a $2k \times 2N$ binary matrix satisfying $\mathbf{D}_{i,j} = 1$ if and only if the i^{th} symbol in the concatenation of messages $\mathbf{x}$ and $\mathbf{y}$ is present on the j^{th} vertical edge between the third and fourth layers. Since the dimension of these vertical edges is at most n, each column of

$\mathbf{D}$ has weight at most n. Thus, there are at least $2k - n$ zeros in each column of $\mathbf{D}$ and, therefore, at least $2N(2k - n)$ zeros in the entire matrix.

Since each sink receives input from only N fourth-layer nodes and must be able to reconstruct all $2k$ components of the messages, every possible choice of N columns must have at least one 1 in each row. Thus, each row in $\mathbf{D}$ must have weight at least $N + 1$, implying that each row in $\mathbf{D}$ has at most $2N - (N + 1) = N - 1$ zeros. Thus, counting along the rows, $\mathbf{D}$ has at most $2k(N-1)$ zeros. Relating this upper bound and the previously calculated lower bound on the number of zeros yields $2N(2k-n) \leq 2k(N-1)$ or equivalently $k/n \leq N/(N+1)$, for arbitrary k and n. Thus, $\epsilon \leq N/(N+1)$.

Now, let $k = N$ and $n = N + 1$, and route the messages as follows:

$$\begin{aligned}
e_{1,2} &= (x_1, \ldots, x_k) \\
e_{1,3} &= (y_1, \ldots, y_k) \\
e_{2,i} &= (x_1, \ldots, x_k) && (4 \leq i \leq 2N + 3) \\
e_{3,i} &= (y_1, \ldots, y_k) && (4 \leq i \leq 2N + 3) \\
e_{i,2N+i} &= (x_1, \ldots, x_k, y_{i-3}) && (4 \leq i \leq N + 3) \\
e_{i,2N+i} &= (y_1, \ldots, y_k, x_{i-(N+3)}) && (N + 4 \leq i \leq 2N + 3).
\end{aligned}$$

Each node in the fourth layer simply passes to its out-edges exactly what it receives on its in-edge. If a sink node in the bottom layer is connected to nodes n_i and n_j where $2N + 4 \leq i \leq 3N + 3$ and $3N + 4 \leq j \leq 4N + 3$ (i.e., a node in the left half of the fourth layer and a node in the right half of the fourth layer) then the sink receives all of message $\mathbf{x}$ from n_i and all of message $\mathbf{y}$ from n_j. On the other hand, if a sink is connected only to nodes in the left half of the fourth layer, then it receives all of message $\mathbf{x}$ from each such node, and receives a distinct component of message $\mathbf{y}$ from each of the fourth-layer nodes, thus giving all of $\mathbf{y}$. A similar situation occurs if a sink node is only connected to fourth-layer nodes on the right half.

Thus, this assignment is a fractional routing solution to $\mathcal{N}_3$. Therefore, $N/(N+1)$ is an achievable routing rate of $\mathcal{N}_3$, so $\epsilon \geq N/(N+1)$. ■

Example 2.3.4. (See Figure 2.4.)

The network $\mathcal{N}_4$ contains a single source n_1 with m messages. The second layer of the network consists of N nodes, each connected to the source via a single edge. The

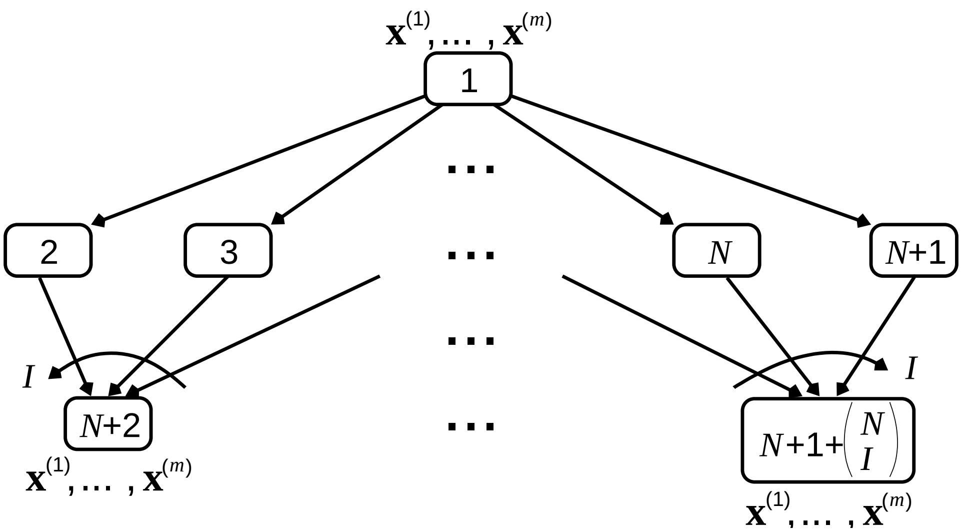

Figure 2.4: The multicast network $\mathcal{N}_4$ whose routing capacity is $N/(m(N-I+1))$.

third layer consists of $\binom{N}{I}$ nodes, each receiving a distinct set of I in-edges from the second layer. Each third-layer node demands all messages. The network is linearly solvable if and only if $m \leq I$ (since the network is multicast and the minimum cut size is I for each sink node [15]). The routing capacity of this network is $\epsilon = N/(m(N-I+1))$.

Proof. In order to meet the demands of each node in the bottom layer, every subset of I nodes in layer two must receive all mk message components from the source. Thus, each of the mk message components must appear at least $N-(I-1)$ times on the N out-edges of the source (otherwise there would be some set of I of the N layer-two nodes not containing some message component). Since the total number of symbols on the N source out-edges is Nn, we must have $mk(N-(I-1)) \leq Nn$ or, equivalently, $k/n \leq N/(m(N-I+1))$, for arbitrary k and n. Hence, $\epsilon \leq N/(m(N-I+1))$.

Now, let $k = N$ and $n = m(N-I+1)$ and denote the components of the m messages (in some order) by $b_1, \ldots, b_{mk}$. Let $\mathbf{D}$ be an $n \times N$ matrix filled with message components from left to right and from top to bottom, with each message component being repeated $N-I+1$ times in a row, i.e., $\mathbf{D}_{i,j} = b_{\lfloor (N(i-1)+j-1)/(N-I+1) \rfloor + 1}$ with $1 \leq i \leq m(N-I+1)$ and $1 \leq j \leq N$.

Let the N columns of the matrix determine the vectors carried on the N out-edges of the source. Since each message component is placed in $N-I+1$ different columns of the matrix, every set of I layer-two nodes will receive all of the mN message components. The $m(N - I + 1) = n$ components at each layer-two node are then transmitted directly to all adjacent layer-three nodes.

Thus, this assignment is a fractional routing solution to $\mathcal{N}_4$. Therefore, $N/(m(N - I + 1))$ is an achievable routing rate of $\mathcal{N}_4$, so $\epsilon \geq N/(m(N - I + 1))$. ■

We next note several facts about the network shown in Figure 2.4.

- The capacity of this network was independently obtained (in a more lengthy argument) by Ngai and Yeung [17]. See also Sanders, Egner, and Tolhuizen [21].
- Ahlswede and Riis [20] studied the case obtained by using the parameters $m = 5, N = 12$, and $I = 8$, which we denote by $\mathcal{N}_5$. They showed that this network has no binary scalar linear solution and yet it has a nonlinear binary scalar solution based upon a $(5, 12, 5)$ Nordstrom-Robinson error correcting code. We note that, by our above calculation, the routing capacity of the Ahlswede-Riis network is $\epsilon = 12/25$.
- Rasala Lehman and Lehman [18] studied the case obtained by using the parameters $m = 2, N = p$, and $I = 2$. They proved that the network is solvable, provided that the alphabet size is at least equal to the square root of the number of sinks. We note that, by our above calculation, the routing capacity of the Rasala Lehman-Lehman network is $\epsilon = p/(2(p - 1))$.
- Using the parameters $m = 2$ and $N = I = 3$ illustrates that the network's routing capacity can be greater than 1. In this case, the network consists of a single source, three second layer nodes, and a single third layer node. The routing capacity of this network is $\epsilon = 3/2$.

Example 2.3.5. (See Figure 2.5.)

This network, due to R. Koetter, was used by Médard et al. [16] to demonstrate that there exists a network with no scalar linear solution but with a vector linear solution. The network consists of two sources, each emitting two messages, and four sinks, each demanding two messages. The network has a vector routing solution of dimension two. The routing capacity of this network is $\epsilon = 1$.

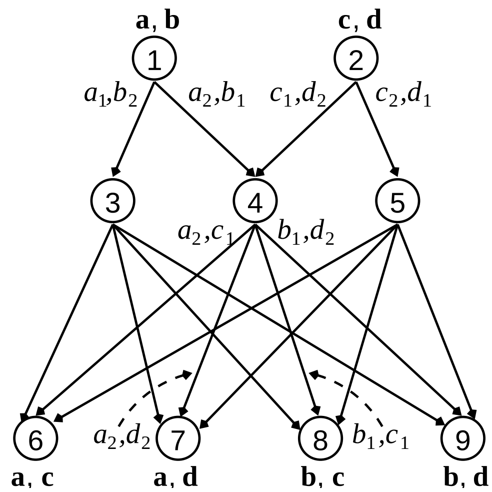

Figure 2.5: The network $\mathcal{N}_6$ whose routing capacity is 1.

Proof. Each source must emit at least $2k$ components and the total capacity of each source's two out-edges is $2n$. Thus, the relation $2k \leq 2n$ must hold, for arbitrary k and n, yielding $\epsilon \leq 1$.

Now let $k = 2$ and $n = 2$, and route the messages as follows (as given in [16]):

$$
\begin{array}{lll}
e_{1,3} = (a_1, b_2) & e_{1,4} = (a_2, b_1) & \\
e_{2,4} = (c_1, d_2) & e_{2,5} = (c_2, d_1) & \\
e_{3,6} = (a_1) & e_{4,6} = (a_2, c_1) & e_{5,6} = (c_2) \\
e_{3,7} = (a_1) & e_{4,7} = (a_2, d_2) & e_{5,7} = (d_1) \\
e_{3,8} = (b_2) & e_{4,8} = (b_1, c_1) & e_{5,8} = (c_2) \\
e_{3,9} = (b_2) & e_{4,9} = (b_1, d_2) & e_{5,9} = (d_1)
\end{array}
$$

This is a fractional routing solution to $\mathcal{N}_6$. Thus, 1 is an achievable routing rate of $\mathcal{N}_6$, so $\epsilon \geq 1$. ■

Example 2.3.6. (See Figure 2.6.)

The network $\mathcal{N}_7$ was demonstrated in [5] to have no linear solution for any vector dimension over a finite field of odd cardinality. The network has three sources n_1, n_2, and n_3 emitting messages $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, respectively. The messages $\mathbf{c}$, $\mathbf{b}$, and $\mathbf{a}$ are demanded by sinks n_{12}, n_{13}, and n_{14}, respectively. The network has no routing solution but does have a coding solution. The routing capacity of this network is $\epsilon = 2/3$.

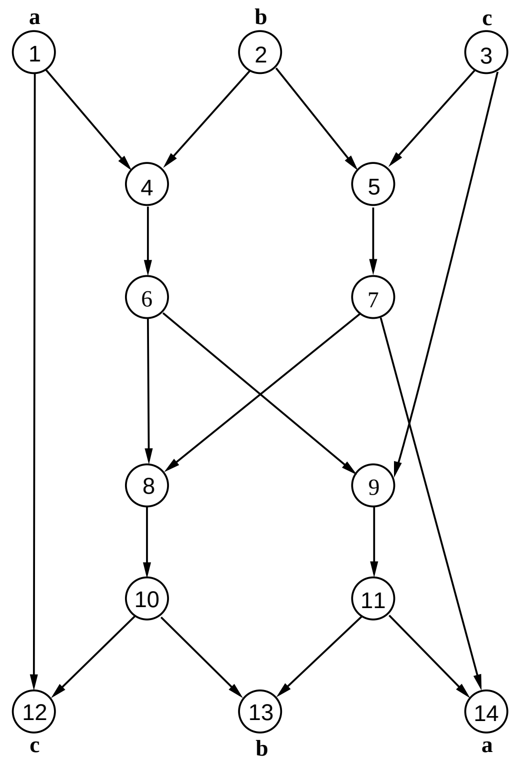

Figure 2.6: The network $\mathcal{N}_7$ whose routing capacity is $2/3$.

Proof. First, note that the edges $e_{1,12}, e_{3,9}$, and $e_{7,14}$ cannot have any affect on a fractional routing solution, so they can be removed. Thus, edges $e_{4,6}$ and $e_{5,7}$ must carry all of the information from the sources to the sinks. Therefore, $3k \leq 2n$, for arbitrary k and n, yielding an upper bound on the routing capacity of $\epsilon \leq 2/3$.

Now, let $k = 2$ and $n = 3$ and route the messages as follows:

$$
\begin{aligned}
e_{1,4} &= (a_1, a_2) & e_{2,4} &= (b_1) \\
e_{2,5} &= (b_2) & e_{3,5} &= (c_1, c_2) \\
e_{4,6} &= (a_1, a_2, b_1) & e_{5,7} &= (c_1, c_2, b_2) \\
e_{6,9} &= (a_1, a_2, b_1) & e_{7,8} &= (b_2, c_1, c_2) \\
e_{8,10} &= (b_2, c_1, c_2) & e_{9,11} &= (a_1, a_2, b_1) \\
e_{10,12} &= (c_1, c_2) & e_{10,13} &= (b_2) \\
e_{11,13} &= (b_1) & e_{11,14} &= (a_1, a_2).
\end{aligned}
$$

This is a fractional routing solution to $\mathcal{N}_7$. Thus, $2/3$ is an achievable routing rate of $\mathcal{N}_7$, so $\epsilon \geq 2/3$. ■

Example 2.3.7. (See Figure 2.7.)

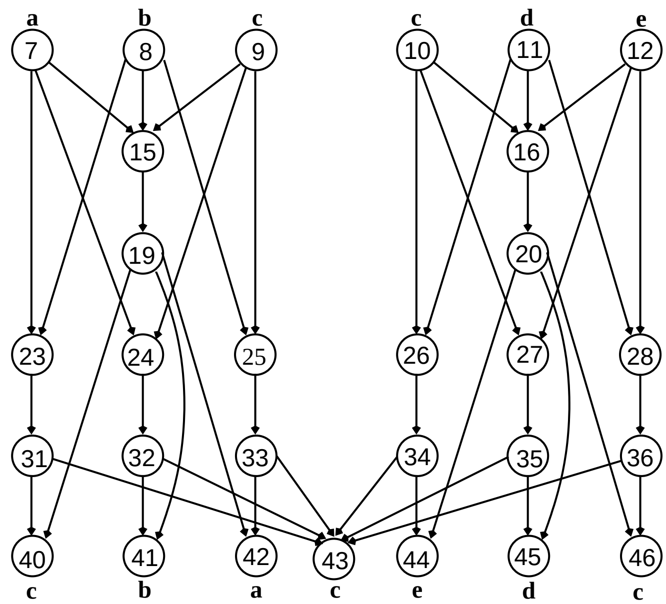

Figure 2.7: The network $\mathcal{N}_8$ whose routing capacity is $1/3$.

The network $\mathcal{N}_8$ shown in Figure 2.7 was given in [5] as a portion of a larger network which was solvable but not vector-linearly solvable. This network piece consists of six sources, n_7 through n_{12}, emitting messages $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{c}, \mathbf{d}$, and $\mathbf{e}$, respectively. The network contains seven sinks, n_{40} through n_{46}, demanding messages $\mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{e}, \mathbf{d}$, and $\mathbf{c}$, respectively. The network has no routing solution but does have a coding solution. The routing capacity of this network is $\epsilon = 1/3$.

Proof. A number of edges in the network do not affect any fractional routing solution and can be removed, yielding the reduced network shown in Figure 2.8. Clearly the demands of node n_{43} are easily met. The remaining portion of the network can be divided into two disjoint, symmetric portions. In each case all $3k$ symbols of information must flow across a single edge (either $e_{15,19}$ or $e_{16,20}$), implying that $3k \leq n$ for arbitrary k and n. Thus, $\epsilon \leq 1/3$.

Now, let $k = 1$ and $n = 3$ and route the messages as follows:

$$e_{15,19} = (a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k)$$

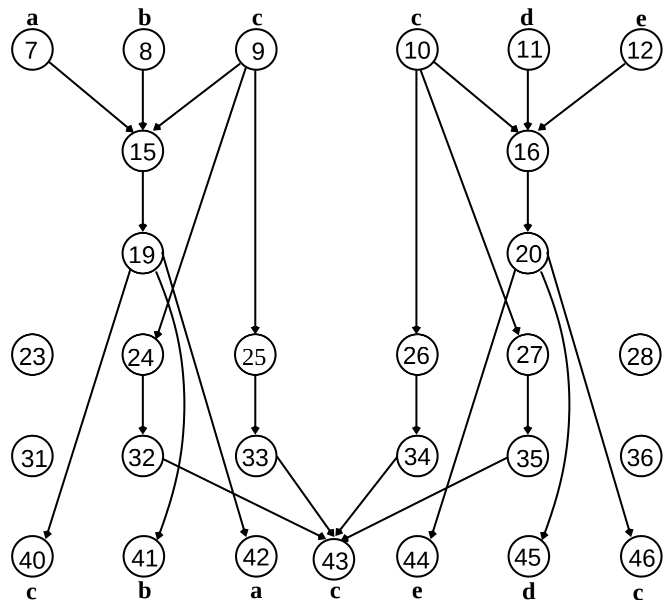

Figure 2.8: Reduced form of the network $\mathcal{N}_8$ given in Figure 2.7.

$$e_{16,20} = (c_1, \ldots, c_k, d_1, \ldots, d_k, e_1, \ldots, e_k).$$

This is a fractional routing solution to $\mathcal{N}_8$. Thus, $1/3$ is an achievable routing rate of $\mathcal{N}_8$, so $\epsilon \geq 1/3$. ■

By combining networks $\mathcal{N}_7$ and $\mathcal{N}_8$ (i.e., by adding shared sources a, b, and c) a network was created which established that linear vector codes are not sufficient for all solvable networks [5]. In the combined network, the two pieces effectively operate independently, and thus the routing capacity of the entire network is limited by the second portion, namely, $\epsilon = 1/3$.

2.4 Routing Capacity Achievability

The examples of the previous section have illustrated various techniques to determine the routing capacity of a network. In this section, some properties of the routing capacity are developed and a concrete method is given, by which the routing capacity of a network can be found.

To begin, a set of inequalities which are satisfied by any minimal fractional routing solution is formulated. These inequalities are then used to prove that the routing capacity of any network is achievable. To facilitate the construction of these inequalities, a variety of subgraphs for a given network are first defined.

Consider a network and its associated graph, $G = (V, E)$, sources S, messages M, and sinks K. For each message $\mathbf{x}$, we say that a directed subgraph of G is an $\mathbf{x}$-*tree* if the subgraph has exactly one directed path from the source emitting $\mathbf{x}$ to each destination node which demands $\mathbf{x}$, and the subgraph is minimal with respect to this property[4]. (Note that such a subgraph can be both an $\mathbf{x}$-tree and a $\mathbf{y}$-tree for distinct messages $\mathbf{x}$ and $\mathbf{y}$.) For each message $\mathbf{x}$, let $s(\mathbf{x})$ denote the number of $\mathbf{x}$-trees. For a given network and for each message $\mathbf{x}$, let $T_1^{\mathbf{x}}, T_2^{\mathbf{x}}, \ldots, T_{s(\mathbf{x})}^{\mathbf{x}}$ be an enumeration of all the $\mathbf{x}$-trees in the network. Figure 2.9 depicts all of the $\mathbf{x}$-trees and $\mathbf{y}$-trees for the network $\mathcal{N}_2$ shown in Figure 2.2.

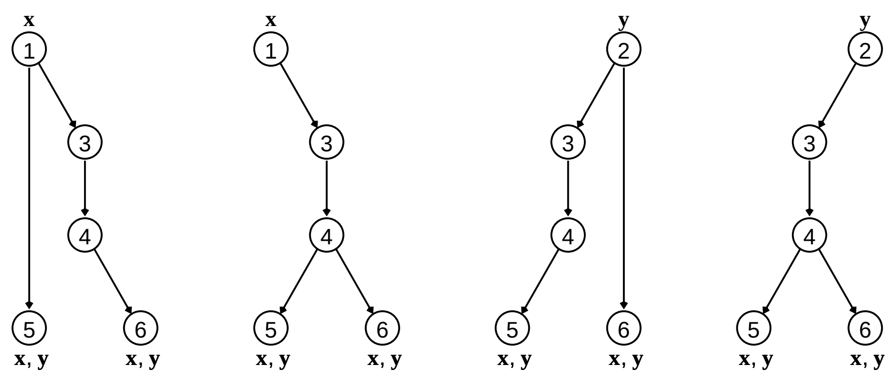

Figure 2.9: All of the $\mathbf{x}$-trees and $\mathbf{y}$-trees of the network $\mathcal{N}_2$.

If $\mathbf{x}$ is a message and j is the unique index in a minimal (k, n) fractional routing solution such that every edge carrying a component x_i appears in $T_j^{\mathbf{x}}$, then we say the $\mathbf{x}$-tree $T_j^{\mathbf{x}}$ *carries* the message component x_i. Such a tree is guaranteed to exist since in the supposed solution each message component must be routed from its source to every destination node demanding the message, and the minimality of the solution ensures that

[4]The definition of an $\mathbf{x}$-tree is similar to that of a directed Steiner tree (also known as a Steiner arborescence). Given a directed, edge-weighted graph, a subset of the nodes in the graph, and a root node, a directed Steiner tree is a minimum-weight subgraph which includes a directed path from the root to every other node in the subset [9]. Thus, an $\mathbf{x}$-tree is a directed Steiner tree where the source node is the root node, the subset contains the source and all sinks demanding $\mathbf{x}$, the edge weights are taken to be 0, and with the additional restrictions that only one directed path from the root to each sink is present, and edges not along these directed paths are not included in the subgraph. In the undirected case, the first additional restriction coupled with the 0-edge-weight case corresponds to the requirement that the subgraph be a tree, which is occasionally incorporated in the definition of a Steiner tree [11].

the edges carrying the message form an $\mathbf{x}$-tree.

Note that we consider $T_i^{\mathbf{x}}$ and $T_j^{\mathbf{y}}$ to be distinct when $\mathbf{x} \neq \mathbf{y}$, even if they are topologically the same directed subgraph of the network. That is, such trees are determined by their topology together with their associated message.

Denote by T_i the i^{th} tree in some fixed ordering of the set

$$\bigcup_{\mathbf{x} \in M} \{T_1^{\mathbf{x}}, \ldots, T_{s(\mathbf{x})}^{\mathbf{x}}\}$$

and define the following index sets:

$$\begin{aligned} A(\mathbf{x}) &= \{i : T_i \text{ is an } \mathbf{x}\text{-tree}\} \\ B(e) &= \{i : T_i \text{ contains edge } e\}. \end{aligned}$$

Note that the sets $A(\mathbf{x})$ and $B(e)$ are determined by the network, rather than by any particular solution to the network. Denote the total number of trees T_i by

$$t = \sum_{\mathbf{x} \in M} s(\mathbf{x}).$$

For any given minimal (k, n) fractional routing solution, and for each $i = 1, \ldots, t$, let c_i denote the number of message components carried by tree T_i in the given solution.

Lemma 2.4.1. *For any given minimal (k, n) fractional routing solution to a nondegenerate network, the following inequalities hold:*

(a) $\displaystyle\sum_{i \in A(\mathbf{x})} c_i \geq k \qquad (\forall \mathbf{x} \in M)$

(b) $\displaystyle\sum_{i \in B(e)} c_i \leq n \qquad (\forall e \in E)$

(c) $0 \leq c_i \leq k \qquad (\forall i \in \{1, \ldots, t\})$

(d) $0 \leq n \leq k|M| \leq kt$.

Proof.

(a) Follows from the fact that all k components of every message must be sent to every destination node demanding them.

(b) Follows from the fact that every edge can carry at most n message components.

(c) Follows from that fact that each message has k components.

(d) Since the routing solution is minimal, it must be the case that $n \leq k|M|$, since edge capacities of size $k|M|$ suffice to carry every component of every message. Also, clearly $|M| \leq t$, since the network is nondegenerate.

■

Lemma 2.4.2. *For any given minimal* (k, n) *fractional routing solution to a nondegenerate network, the following inequalities, over the real variables* $d_1, \ldots, d_t, \rho$*, have a rational solution*[5]*:*

$$\sum_{i \in A(\mathbf{x})} d_i \geq 1 \qquad (\forall \mathbf{x} \in M) \tag{2.1}$$

$$\sum_{i \in B(e)} d_i \leq \rho \qquad (\forall e \in E) \tag{2.2}$$

$$0 \leq d_i \leq 1 \qquad (\forall i \in \{1, \ldots, t\}) \tag{2.3}$$

$$0 \leq \rho \leq t \tag{2.4}$$

by choosing $d_i = c_i/k$ *and* $\rho = n/k$*.*

Proof. Inequalities (2.1)–(2.4) follow immediately from Lemma 2.4.1(a)–(d), respectively, by division by k.

■

We refer to (2.1)–(2.4) as the *network inequalities* associated with a given network.[6] Note that the routing rate in the given (k, n) fractional routing solution in Lemma 2.4.2 is $1/\rho$.

[5] If a solution $(d_1, \ldots, d_t, \rho)$ to these inequalities has all rational components, then it is said to be a *rational solution*.

[6] Similar inequalities are well-known for undirected network flow problems (e.g., see [11] for the case of single-source networks).

For convenience, define the sets

$$\begin{aligned} V &= \{\rho \in \mathbb{R} : (d_1, \ldots, d_t, \rho) \text{ is a solution to the} \\ &\qquad \text{network inequalities for some } (d_1, \ldots, d_t)\} \\ \hat{V} &= \{r : 1/r \in V\}. \end{aligned}$$

Lemma 2.4.3. *If the network inequalities corresponding to a nondegenerate network have a rational solution with $\rho > 0$, then there exists a fractional routing solution to the network with achievable routing rate $1/\rho$.*

Proof. Let $(d_1, \ldots, d_t, \rho)$ be a rational solution to the network inequalities with $\rho > 0$. To construct a fractional routing solution, let the dimension k of the messages be equal to the least common multiple of the denominators of the non-zero components of $(d_1, \ldots, d_t, \rho)$. Also, let the capacity of the edges be $n = k\rho$, which is an integer. Now, for each $i = 1, \ldots, t$, let $c_i = d_i k$, each of which is an integer. A (k, n) fractional routing solution can be constructed by, for each message $\mathbf{x}$, arbitrarily partitioning the k components of the message over all $\mathbf{x}$-trees such that exactly c_i components are sent along each associated tree T_i. ∎

The following corollary shows that the set U (defined in Section 2.2) of achievable routing rates of any network is the same as the set of reciprocals of rational ρ that satisfy the corresponding network inequalities.

Corollary 2.4.4. *For any nondegenerate network, $\hat{V} \cap \mathbb{Q} = U$.*

Proof. Lemma 2.4.2 implies that $U \subseteq \hat{V} \cap \mathbb{Q}$ and Lemma 2.4.3 implies that $\hat{V} \cap \mathbb{Q} \subseteq U$. ∎

We next use the network inequalities to prove that the routing capacity of a network is achievable. To prove this property, the network inequalities are viewed as a set of inequalities in $t + 1$ variables, $d_1, \ldots, d_t, \rho$, which one can attempt to solve. By formulating a linear programming problem, it is possible to determine a fractional routing solution to the network which achieves the routing capacity. As a consequence, the routing capacity of every network is rational and the routing capacity of every nondegenerate network is achievable. The following theorem gives the latter result in more detail.

Theorem 2.4.5. *The routing capacity of every nondegenerate network is achievable.*

Proof. We first demonstrate that the network inequalities can be used to determine the routing capacity of a network. Let

$$H = \{(d_1, \ldots, d_t, \rho) \in \mathbb{R}^{t+1} : \text{ the network inequalities are satisfied}\}$$
$$\rho_0 = \inf V$$

and define the linear function

$$f(d_1, \ldots, d_t, \rho) = \rho.$$

Note that H is non-empty since a rational solution to the network inequalities can be found for any network by setting $d_i = 1, \ \forall i$ and $\rho = t$. Also, since H is compact (i.e., a closed and bounded polytope), the restriction of f to H achieves its infimum ρ_0 on H. Thus, there exist $\hat{d}_1, \ldots, \hat{d}_t \in \mathbb{R}$ such that $(\hat{d}_1, \ldots, \hat{d}_t, \rho_0) \in H$. In fact, a linear program can be used to minimize f on H, yielding ρ_0. Furthermore, since the variables $d_1, \ldots, d_t, \rho$ in the network inequalities have rational coefficients, we can assume without loss of generality that $\hat{d}_1, \ldots, \hat{d}_t, \rho_0 \in \mathbb{Q}$. Now, by Corollary 2.4.4, we have

$$\begin{aligned}
\epsilon &= \sup U \\
&= \sup \left(\hat{V} \cap \mathbb{Q}\right) \\
&= \sup\{r \in \mathbb{Q} : (d_1, \ldots, d_t, 1/r) \in H\} \\
&= \sup\{1/\rho \in \mathbb{Q} : (d_1, \ldots, d_t, \rho) \in H\} \\
&= \max\{1/\rho \in \mathbb{Q} : (d_1, \ldots, d_t, \rho) \in H\} \\
&= 1/\rho_0.
\end{aligned}$$

Thus, the network inequalities can be used to determine the routing capacity of a network.

Furthermore, the fractional routing solution induced by the solution $(\hat{d}_1, \ldots, \hat{d}_t, \rho_0)$ to the network inequalities has achievable routing rate $1/\rho_0 = \epsilon$. Thus, the routing capacity of any network is achievable.

■

Corollary 2.4.6. *The routing capacity of every network is rational.*

Proof. If a network is degenerate, then its capacity is zero, which is rational. Otherwise, Theorem 2.4.5 guarantees that there exists a (k, n) fractional routing solution such that the routing capacity equals k/n, which is rational. ■

Since any linear programming algorithm (e.g., the simplex method) will work in the proof of Theorem 2.4.5, we also obtain the following corollary.

Corollary 2.4.7. *There exists an algorithm for determining the routing capacity of a network.*

We note that the results in Section 2.4 can be generalized to networks whose edge capacities are arbitrary rational numbers. In such case, the term ρ in (2.2) of the network inequalities would be multiplied by the capacity of the edge e, and the term t in (2.4) would be multiplied by the maximum edge capacity.

2.5 Network Construction for Specified Routing Capacity

Given any rational number $r \geq 0$, it is possible to form a network whose routing capacity is $\epsilon = r$. The following two theorems demonstrate how to construct such networks. The first theorem considers the general case when $r \geq 0$, but the resulting network is unsolvable (i.e., for $k = n$) for $r < 1$. The second theorem considers the case when $0 < r \leq 1$ and yields a solvable network.

Theorem 2.5.1. *For each rational $r \geq 0$, there exists a network whose routing capacity is $\epsilon = r$.*

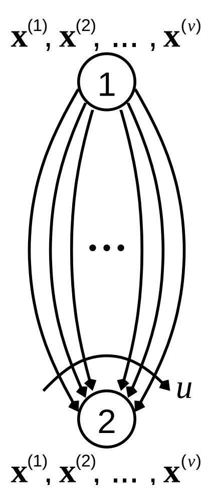

Figure 2.10: A network $\mathcal{N}_9$ that has routing capacity $r = u/v \geq 0$.

Proof. If $r = 0$ then any degenerate network suffices. Thus, assume $r > 0$ and let $r = u/v$ where u and v are positive integers. Consider a network with a single source and a single sink connected by u edges, as shown in Figure 2.10. The source emits messages $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(v)}$ and all messages are demanded by the sink. Let k denote the message dimension and n denote the edge capacity.

In a fractional routing solution, the full vk components must be transferred along the u edges of capacity n. Thus, for a fractional routing solution to exist, we require $vk \leq un$, and hence the routing capacity is upper bounded by u/v.

If $k = u$ and $n = v$, then $kv = uv$ message components can be sent arbitrarily along the u edges since the cumulative capacity of all the edges is $nu = vu$. Thus, the routing capacity upper bound is achievable.

Thus, for each rational $r \geq 0$, a single-source, single-sink network can be constructed which has routing capacity $\epsilon = r$.

■

The network $\mathcal{N}_9$ discussed in Theorem 2.5.1 is unsolvable for $0 < r < 1$, since the min cut across the network does not have the required transmission capacity. However, the network is indeed solvable for $r \geq 1$ using a routing solution.

Theorem 2.5.2. *For each rational $r \in (0, 1]$ there exists a solvable network whose routing capacity is $\epsilon = r$.*

Proof. Let $r = p/m$ where $p \leq m$. Consider a network with four layers, as shown in Figure 2.11 where all edges point downward. The network contains m sources, all in the first layer. Each source emits a unique message, yielding messages $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ in the network. The second layer of the network contains p nodes, each of which is connected to all m sources, forming a complete connection between the first and second layers. The third layer also contains p nodes and each is connected in a straight through fashion to a corresponding node in the second layer. The fourth layer consists of m sinks, each demanding all m messages. The third and fourth layers are also completely connected. Finally, each sink is connected to a unique set of $m - 1$ sources, forming a complete connection except the straight through edges between the first and fourth layers. Thus, the network can be thought of as containing both a direct and an indirect route between the sources and sinks.

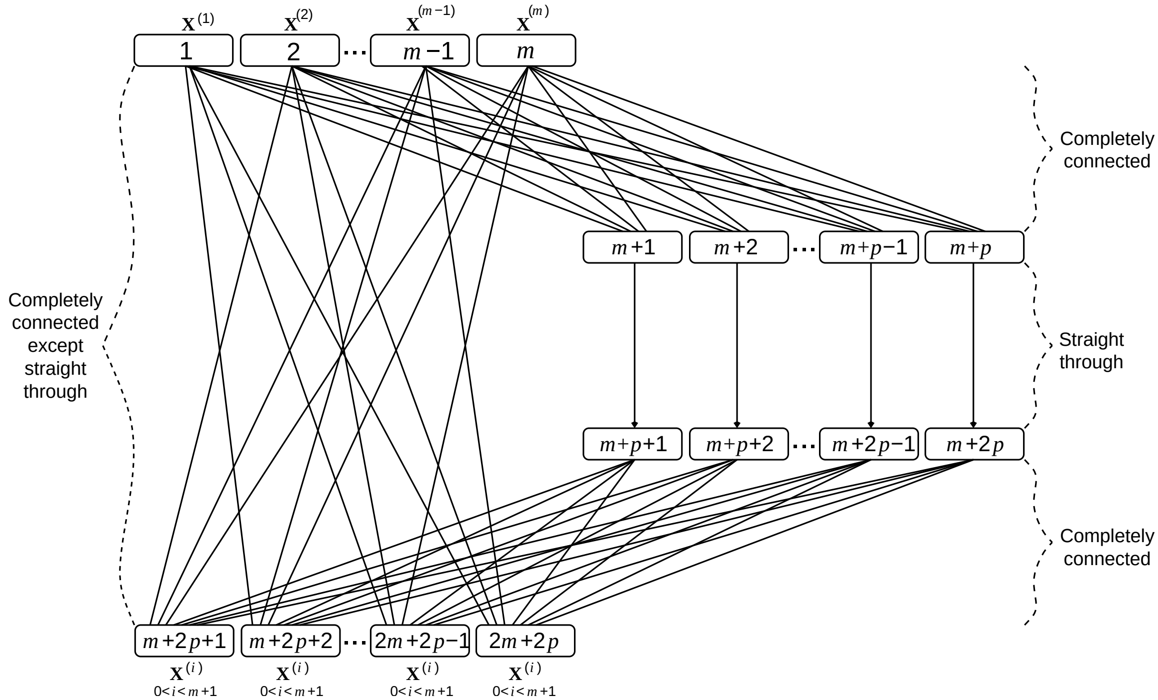

Figure 2.11: A solvable network $\mathcal{N}_{10}$ that has routing capacity $r = p/m \in (0, 1]$. All edges in the network point downward.

The routing capacity of this network is now shown to be $\epsilon = r = p/m$. Let k be the dimension of the messages and let n be the capacity of the edges. To begin, the routing capacity is demonstrated to be upper bounded by p/m. First, note that since each sink is directly connected to all but one of the sources and since $r = p/m \leq 1$, each sink can receive all but one of the messages directly. Furthermore, in each case, the missing message must be transmitted to the sink along the indirect route (from the source through the second and third layers to the sink). Since each of the m messages is missing from one of the sinks, a total of mk message components must be transmitted along the indirect paths. The cumulative capacity of the indirect paths is pn, as clearly seen by considering the straight through connections between layers two and three. Thus, the relation $mk \leq pn$ must hold, yielding $k/n \leq p/m$ for arbitrary k and n. Thus $\epsilon \leq p/m$.

To prove that this upper bound on the routing capacity is achievable, consider a solution which sets $k = p$ and $n = m$. As noted previously, direct transmission of $m - 1$ of the messages to each sink is clearly possible. Now, each second-layer node receives all k components of all m messages, for a total of $mk = mp$ components. The cumulative capacity of the links from the second to third layers is $pn = pm$. Thus, since the sinks receive all data received by the third-layer nodes, the mp message components can be assigned arbitrarily to the pm straight-through slots, allowing each sink to receive the correct missing message. Hence, this assignment is a fractional routing solution. Therefore, p/m is an achievable routing rate of the network, so $\epsilon \geq p/m$.

Now, the network is shown to be solvable by presenting a solution. Let the alphabet from which the components of the messages are drawn be an Abelian group. As previously, all but one message is received by each source along the direct links from the sources to the sinks. Now, note that node n_{m+1} receives all m messages from the sources. Thus, it is possible to send the combination $\mathbf{x}^{(1)} + \mathbf{x}^{(2)} + \cdots + \mathbf{x}^{(m)}$ along edge $e_{m+1,m+p+1}$. Node n_{m+p+1} then passes this combination along to each of the sinks. Since each sink possesses all but one message, it can extract the missing message from the combination received from node n_{m+p+1}. Thus, the demands of each sink are met.

Hence, the generalized network shown in Figure 2.11 represents a solvable network whose routing capacity is the rational $r = p/m \in (0, 1]$.

■

In the network $\mathcal{N}_{10}$, a routing solution (with $k = n$) would require all m messages to be transmitted along the p straight-through paths in the indirect portion of the network. However, for $r \in (0, 1)$ we have $p < m$, hence no routing solution exists. Thus, the network requires coding to achieve a solution. Also, note that if the network $\mathcal{N}_{10}$ is specialized to the case $m = 2$ and $p = 1$, then it becomes the network in Figure 2.2.

2.6 Coding Capacity

This section briefly considers the coding capacity of a network, which is a generalization of the routing capacity. The coding capacity is first defined and two examples are then discussed. Finally, it is shown that the coding capacity is independent of the chosen alphabet.

A (k, n) *fractional coding solution* of a network is a coding solution that uses messages with k components and edges with capacity n. If a network has a (k, n) fractional coding solution, then the rational number k/n is said to be an *achievable coding rate*. The *coding capacity* is then given by

$$\gamma = \sup \{r \in \mathbb{Q} : r \text{ is an achievable coding rate}\} .$$

If a (k, n) fractional coding solution uses only linear coding, then k/n is an *achievable linear coding rate* and we define the *linear coding capacity* to be

$$\lambda = \sup \{r \in \mathbb{Q} : r \text{ is an achievable linear coding rate}\} .$$

Note that unlike fractional routing solutions, fractional coding solutions must be considered in the context of a specific alphabet. Indeed, the linear coding capacity in general depends on the alphabet [5]. However, it will be shown in Theorem 2.6.5 that the coding capacity of a network is independent of the chosen alphabet.

Clearly, for a given alphabet, the coding capacity of a network is always greater than or equal to the linear coding capacity. Also, if a network is solvable (i.e., with $k = n$), then the coding capacity is greater than or equal to 1, since $k/n = k/k$ is an achievable coding rate. Similarly, if a network is linearly solvable, then the linear coding capacity is greater than or equal to 1.

The following examples illustrate the difference between the routing capacity and coding capacity of a network.

Example 2.6.1. The special case $\mathcal{N}_5$ of the network shown in Figure 2.4 has routing capacity $\epsilon = 12/25$, as discussed in the note following Example 2.3.4. Using a cut argument, it is clear that the coding capacity of the network is upper bounded by $8/5$, since each sink demands $5k$ message components and has a total capacity of $8n$ on its incoming edges. Lemmas 2.6.2 and 2.6.3 will respectively prove that this network has a scalar linear solution for every finite field other than $GF(2)$ and has a vector linear solution for $GF(2)$. Consequently, the linear coding capacity for any finite field alphabet is at least 1, which is strictly greater than the routing capacity.

Lemma 2.6.2. *Network $\mathcal{N}_5$ has a scalar linear solution for every finite field alphabet other than $GF(2)$.*

Proof. Let $a, b, c, d,$ and e be the messages at the source. Let the alphabet be a finite field F with $|F| > 2$. Let $z \in F - \{0, 1\}$. Define the following sets (D is a multiset):

$$A = \{a, b, c, d, e\}$$

$$\begin{aligned} B &= \{za+b, zb+c, zc+d, zd+e, ze+a\} \\ C &= \{a+b+c+d+e\} \\ D &= A \cup B \cup C \cup C. \end{aligned}$$

Then $|D| = 12$. Let the symbols carried on the 12 edges emanating from the source correspond to a specific permutation of the 12 elements of D. We will show that the demands of all $\binom{12}{8}$ sinks are satisfied by showing that all of the messages $a, b, c, d,$ and e can be recovered (linearly) from every multiset $S \subset D$ satisfying $|S| = 8$.

If $|S \cap A| = 5$ then the recovery is trivial.

If $|S \cap A| = 4$ then without loss of generality assume $e \notin S$. If $a+b+c+d+e \in S$, then e can clearly be recovered. If $a+b+c+d+e \notin S$, then $|S \cap B| = 4$, in which case $\{zd+e, ze+a\} \cap S \neq \emptyset$, and thus e can be recovered.

If $|S \cap A| = 1$ then $B \subset S$, so the remaining 4 elements of A can be recovered.

If $|S \cap A| = 2$ then $|B \cap S| \geq 4$, so the remaining 3 elements of A can be recovered.

If $|S \cap A| = 3$ then $|B \cap S| \geq 3$. If $|B \cap S| \geq 4$, then the remaining 2 elements of A can be recovered, so assume $|B \cap S| = 3$, in which case $a+b+c+d+e \in S$. Due to the symmetries of the elements in B, we assume without loss of generality that $A \cap S \in \{\{a,b,c\}, \{a,b,d\}\}$. First consider the case when $A \cap S = \{a,b,c\}$. Then, $d+e$ can be recovered. If $zd+e \in S$ then we can solve for d and e since $z \neq 1$. If $zd+e \notin S$ then $S \cap \{zc+d, ze+a\} \neq \emptyset$, so either d can be recovered from c and $zc+d$ or e can be recovered from a and $ze+a$. Then the remaining term is recoverable from $d+e$. Now consider the case when $A \cap S = \{a,b,d\}$. Then $c+e$ can be recovered. If $S \cap \{zb+c, zc+d\} \neq \emptyset$ then c can be recovered from either b and $zb+c$ or d and $zc+d$. If $S \cap \{zb+c, zc+d\} = \emptyset$ then $S \cap \{zd+e, ze+a\} \neq \emptyset$, so e can be recovered from either d and $zd+e$ or a and $ze+a$. Finally, the remaining term can be recovered from $c+e$.

■

Lemma 2.6.3. *Network $\mathcal{N}_5$ has a binary linear solution for vector dimension* 2.

Proof. Consider a scalar linear solution over $GF(4)$ (which is known to exist by Lemma 2.6.2). The elements of $GF(4)$ can be viewed as the following four 2×2 matrices over

$GF(2)$:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Then, using the $GF(4)$ solution from Lemma 2.6.2 and substituting in the matrix representation yields the following 12 linear functions of dimension 2 for the second layer of the network:

$$\begin{gathered}
\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \\
\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \\
\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \\
\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \\
\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} - \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \\
\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \\
\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \\
\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.
\end{gathered}$$

It is straightforward to verify that from any 8 of these 12 vector linear functions, one can linearly obtain the 5 message vectors $\binom{a_1}{a_2}, \binom{b_1}{b_2}, \binom{c_1}{c_2}, \binom{d_1}{d_2}, \binom{e_1}{e_2}$.

■

Example 2.6.4. As considered in Example 2.3.1, the network $\mathcal{N}_1$ has routing capacity $\epsilon = 3/4$. We now show that both the coding and linear coding capacities are equal to 1, which is strictly greater than the routing capacity.

Proof. Network $\mathcal{N}_1$ has a well known scalar linear solution [1] given by

$$e_{1,2} = e_{2,4} = e_{2,6} = x$$

$$e_{1,3} = e_{3,4} = e_{3,7} = y$$
$$e_{4,5} = e_{5,6} = e_{5,7} = x + y.$$

Thus, $\lambda \geq 1$ and $\gamma \geq 1$.

To upper bound the coding and linear coding capacities, note that each sink demands both messages but only possesses two incoming edges. Thus, we have the requirement $2k \leq 2n$, for arbitrary k and n. Hence, $\lambda \leq 1$ and $\gamma \leq 1$.

■

Theorem 2.6.5. *The coding capacity of any network is independent of the alphabet used.*

Proof. Suppose a network has a (k, n) fractional coding solution over an alphabet A and let B be any other alphabet of cardinality at least two. Let $\epsilon > 0$ and let

$$t = \left\lceil \frac{(k+1)\log_2 |B|}{n\epsilon \log_2 |A|} \right\rceil .$$

There is clearly a (tk, tn) fractional coding solution over the alphabet A obtained by independently applying the (k, n) solution t times. Define the quantities

$$n' = n \left\lceil t \cdot \frac{\log_2 |A|}{\log_2 |B|} \right\rceil$$
$$k' = \left\lfloor \frac{kn'}{n} \right\rfloor - k$$

and notice by some computation that

$$|B|^{n'} \geq |A|^{tn} \tag{2.5}$$
$$|B|^{k'} \leq |A|^{tk} \tag{2.6}$$
$$\frac{k'}{n'} \geq \frac{k}{n} - \epsilon. \tag{2.7}$$

For each edge e, let d_e and m_e respectively be the number of relevant in-edges and messages originating at the starting node of e, and, for each node v let d_v and m_v respectively be the number of relevant in-edges and messages originating at v. For each edge e, denote the edge encoding function for e by

$$f_e : (A^{tn})^{d_e} \times (A^{tk})^{m_e} \to A^{tn}$$

and for each node v, and each message $\mathbf{m}$ demanded by v denote the corresponding node decoding function by

$$f_{v,\mathbf{m}} : (A^{tn})^{d_v} \times (A^{tk})^{m_v} \to A^{tk}.$$

The function f_e determines the vector carried on the out-edge e of a node based upon the vectors carried on the in-edges and the message vectors originating at the same node. The function $f_{v,\mathbf{m}}$ attempts to produce the message vector $\mathbf{m}$ as a function of the vectors carried on the in-edges of the node v and the message vectors originating at v. Let $h : A^{tn} \to B^{n'}$ and $h_0 : B^{k'} \to A^{tk}$ be any injections (they exist by (2.5) and (2.6)). Define $\hat{h} : B^{n'} \to A^{tn}$ such that $\hat{h}(h(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in A^{tn}$ and $\hat{h}(\mathbf{x})$ is arbitrary otherwise. Also, define $\hat{h}_0 : A^{tk} \to B^{k'}$ such that $\hat{h}_0(h_0(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in B^{k'}$ and $\hat{h}_0(\mathbf{x})$ is arbitrary otherwise. Define for each edge e the mapping

$$g_e : (B^{n'})^{d_e} \times (B^{k'})^{m_e} \to B^{n'}$$

by

$$g_e(\mathbf{x}_1, \ldots, \mathbf{x}_{d_e}, \mathbf{y}_1, \ldots, \mathbf{y}_{m_e}) \\ = h(f_e(\hat{h}(\mathbf{x}_1), \ldots, \hat{h}(\mathbf{x}_{d_e}), h_0(\mathbf{y}_1), \ldots, h_0(\mathbf{y}_{m_e})))$$

for all $\mathbf{x}_1, \ldots, \mathbf{x}_{d_e} \in B^{n'}$ and for all $\mathbf{y}_1, \ldots, \mathbf{y}_{m_e} \in B^{k'}$. Similarly, define for each node v and each message $\mathbf{m}$ demanded at v the mapping

$$g_{v,\mathbf{m}} : (B^{n'})^{d_v} \times (B^{k'})^{m_v} \to B^{k'}$$

by

$$g_{v,\mathbf{m}}(\mathbf{x}_1, \ldots, \mathbf{x}_{d_v}, \mathbf{y}_1, \ldots, \mathbf{y}_{m_v}) \\ = \hat{h}_0(f_{v,\mathbf{m}}(\hat{h}(\mathbf{x}_1), \ldots, \hat{h}(\mathbf{x}_{d_v}), h_0(\mathbf{y}_1), \ldots, h_0(\mathbf{y}_{m_v})))$$

for all $\mathbf{x}_1, \ldots, \mathbf{x}_{d_v} \in B^{n'}$ and for all $\mathbf{y}_1, \ldots, \mathbf{y}_{m_v} \in B^{k'}$.

Now consider the (k', n') fractional network code over the alphabet B obtained by using the edge functions g_e and decoding functions $g_{v,\mathbf{m}}$. For each edge in the network, the vector carried on the edge in the (k, n) solution over the alphabet A and the vector carried on the edge in the (k', n') fractional network code over B can each be obtained from the other using h and $\hat{h}$, and likewise for the vectors obtained at sink nodes from the decoding functions for the alphabets A and B (using h_0 and $\hat{h}_0$). Thus, the set of edge functions g_e and decoding functions $g_{v,\mathbf{m}}$ gives a (k', n') fractional routing solution of the network over alphabet B, since the vector on every edge in the solution over A can be determined (using h, h_0, $\hat{h}$, and $\hat{h}_0$) from the vector on the same edge in the solution over B. The (k', n') solution achieves a rate of k'/n', which by (2.7) is at least $(k/n) - \epsilon$. Since ϵ was chosen as an arbitrary positive number, the supremum of achievable rates of the network over the alphabet B is at least k/n. Thus, if a coding rate is achievable by one alphabet, then that rate is a lower bound to the coding capacity for all alphabets. This implies the network coding capacity (the supremum of achievable rates) is the same for all alphabets. ■

There are numerous interesting open questions regarding coding capacity, some of which we now mention. Is the coding capacity (resp. linear coding capacity) achievable and/or rational for every network? For which networks is the linear coding capacity smaller than the coding capacity, and for which networks is the routing capacity smaller than the linear coding capacity? Do there exist algorithms for computing the coding capacity and linear coding capacity of networks?

2.7 Conclusions

This paper formally defined the concept of the routing capacity of a network and proved a variety of related properties. When fractional routing is used to solve a network, the dimension of the messages need not be the same as the capacity of the edges. The routing capacity provides an indication of the largest possible fractional usage of the edges for which a fractional routing solution exits. A variety of sample networks were considered to illustrate the notion of the routing capacity. Through a constructive procedure, the routing capacity of any network was shown to be achievable and rational. Furthermore, it was demonstrated that every rational number in $(0, 1]$ is the routing capacity of some solvable network. Finally, the coding capacity of a network was also defined and was proven to be independent of the alphabet used.

The results in this paper straightforwardly generalize to (not necessarily acyclic) undirected networks and to directed networks with cycles as well. Also, the results can be generalized to networks with nonuniform (but rational) edge capacities; in such case, some extra coefficients are required in the network inequalities. An interesting future problem would be to find a more efficient algorithm for computing the routing capacity of a network.

2.8 Acknowledgment

The authors thank Emina Soljanin and Raymond Yeung for providing helpful references.

The text of this chapter, in full, is a reprint of the material as it appears in Jillian Cannons, Randall Dougherty, Chris Freiling, and Kenneth Zeger, "Network Routing Capacity," *IEEE Transactions on Information Theory*, vol. 52, no. 3, March 2006.

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.

[2] C. Chekuri, C. Fragouli, and E. Soljanin "On average throughput benefits and alphabet size in network coding," *IEEE Trans. Inf. Theory*, submitted for publication.

[3] P. A. Chou, Y. Wu, and K. Jain, "Practical network coding," *Proc. Allerton Conf. Communication, Control, and Computing*, Oct. 2003.

[4] R. Dougherty, C. Freiling, and K. Zeger, "Linearity and solvability in multicast networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2243–2256, Oct. 2004.

[5] R. Dougherty, C. Freiling, and K. Zeger, "Insufficiency of linear coding in network information flow," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2745–2759, Aug. 2005.

[6] R. Dougherty, C. Freiling, and K. Zeger, "Unachievability of network coding capacity", *IEEE Trans. Inf. Theory*, (Joint Special Issue with *IEEE/ACM Trans. Netw.*, to be published.

[7] M. Feder, D. Ron, and A. Tavory, "Bounds on linear codes for network multicast," in *Electronic Colloquium on Computational Complexity (ECCC)*, 2003, Rep. 33, pp. 1–9.

[8] T. Ho, D. Karger, M. Médard, and R. Koetter, "Network coding from a network flow perspective," in *Proc. IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul., p. 441.

[9] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem, Volume 53 of the Annals of Discrete Mathematics*. Amsterdam, The Netherlands: Elsevier Science Publishers (North-Holland), 1992.

[10] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. Tolhuizen, "Polynomial time algorithms for multicast network code construction," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1973–1982, Jun. 2005.

[11] K. Jain, M. Mahdian, and M. Salavatipour, "Packing Steiner trees," in *Proc. 14th Annual ACM-SIAM Symp. Discrete Algorithms*, Baltimore, MD, Jan. 2003, pp. 266–274.

[12] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 782–795, Oct. 2003.

[13] Z. Li and B. Li, "Network coding in undirected networks," in *Proc. 38th Annu. Conf. Information Sciences and Systems (CISS)*, Princeton, NJ, Mar. 2004. Paper #603.

[14] Z. Li, B. Li, D. Jiang, and L. C. Lau. (2004, Feb.) "On achieving optimal end-to-end throughput in data networks: Theoretical and empirical studies," ECE Tech. Rep., Dep. Elec. Comp. Eng., Univ. Toronto, Toronto, ON, Canada. [Online]. Available: http://www.eecg.toronto.edu/∼bli/research.html

[15] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371–381, Feb. 2003.

[16] M. Médard, M. Effros, T. Ho, D. Karger, "On coding for non-multicast networks," in *Proc. 41st Annu. Allerton Conf. Communication Control and Computing*, Monticello, IL, Oct. 2003.

[17] C. K. Ngai and R. W. Yeung, "Network coding gain of combination networks," in *Proc. IEEE Information Theory Workshop (ITW)*, San Antonio, TX, Jan. 2005.

[18] A. Rasala Lehman and E. Lehman, "Complexity classification of network information flow problems," in *Proc. 41st Annu. Allerton Conf. Communication Control and Computing*, Monticello, IL, Oct. 2003.

[19] A. Rasala Lehman and E. Lehman, "Network information flow: Does the model need tuning?," in *Proc. Symp. Discrete Algorithms (SODA)*, Vancouver, BC, Canada, Jan. 2005, pp. 499–504.

[20] S. Riis, "Linear versus non-linear boolean functions in network flow," in *Proc. 38th Annu. Conf. Information Sciences and Systems (CISS)*, Princeton, NJ, March 2004.

[21] P. Sanders, S. Egner, L. Tolhuizen, "Polynomial time algorithms for network information flow", in *Proc. 15th ACM Symp. Parallelism in Algorithms and Architectures (SPAA)*, San Diego, CA, Jun. 2003, pp. 286–294.

[22] F. Shahrokhi and D. W. Matula, “The maximum concurrent flow problem,” *J. ACM*, vol. 37, no. 2, pp. 318–334, 1990.

[23] Y. Wu, P. A. Chou, and K. Jain, “A comparison of network coding and tree packing,” in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Chicago, IL, Jun./Jul. 2004, p. 145.

[24] R. W. Yeung, *A First Course in Information Theory*. New York: Kluwer, 2002.

Chapter 3

Network Coding Capacity With a Constrained Number of Coding Nodes

Abstract

We study network coding capacity under a constraint on the total number of network nodes that can perform coding. That is, only a certain number of network nodes can produce coded outputs, whereas the remaining nodes are limited to performing routing. We prove that every non-negative, monotonically non-decreasing, eventually constant, rational-valued function on the non-negative integers is equal to the capacity as a function of the number of allowable coding nodes of some directed acyclic network.

3.1 Introduction

Let $\mathbb{N}$ denote the positive integers, and let $\mathbb{R}$ and $\mathbb{Q}$ denote the real and rational numbers, respectively, with a superscript "$+$" denoting restriction to positive values. In this paper, a *network* is a directed acyclic multigraph $G = (V, E)$, some of whose nodes are information sources or receivers (e.g. see [13]). Associated with the sources are m generated *messages*, where the i^{th} source message is assumed to be a vector of k_i arbitrary elements of a fixed finite alphabet $\mathcal{A}$ of size at least two. At any node in the network, each out-edge carries a vector of n alphabet symbols which is a function (called an *edge function*) of the vectors of symbols carried on the in-edges to the node, and of the node's message vectors if it is a source. Each network edge is allowed to be used at most once (thus, at most n symbols can travel across each edge). It is assumed that every network edge is reachable by some source message. Associated with each receiver are *demands*, which are subsets of the network messages. Each receiver has *decoding functions* which map the receiver's inputs to vectors of symbols in an attempt to produce the messages demanded at the receiver. The goal is for each receiver to deduce its demanded messages from its in-edges and source messages by having information propagate from the sources through the network.

A $(k_1, \ldots, k_m, n)$ *fractional code* is a collection of edge functions, one for each edge in the network, and decoding functions, one for each demand of each receiver in the network. A $(k_1, \ldots, k_m, n)$ *fractional solution* is a $(k_1, \ldots, k_m, n)$ fractional code which results in every receiver being able to compute its demands via its decoding functions, for all possible assignments of length-k_i vectors over the alphabet to the i^{th} source message, for all i. An edge function performs routing when it copies specified input components to its output components. A node performs *routing* when the edge function of each of its out-edges performs routing. Whenever an edge function for an out-edge of a node depends only on the symbols of a single in-edge of that node, we assume, without loss of generality, that the out-edge carries the same vector of symbols as the in-edge it depends on.

For each i, the ratio k_i/n can be thought of as the rate at which source i injects data into the network. Thus, different sources can produce data at different rates. If a network has a $(k_1, \ldots, k_m, n)$ fractional solution over some alphabet, then we say that $(k_1/n, \ldots, k_m/n)$ is an *achievable rate vector*, and we define the *achievable rate region*[1] of the network as the set

$$S = \{r \in \mathbb{Q}^m : r \text{ is an achievable rate vector}\}.$$

Determining the achievable rate region of an arbitrary network appears to be a formidable task. Consequently, one typically studies certain scalar quantities called coding capacities, which are related to achievable rates. A routing capacity of a network is a coding capacity

[1] Sometimes in the literature the closure $\bar{S}$, with respect to $\mathbb{R}^m$, is taken as the definition of the achievable rate region.

under the constraint that only routing is permitted at network nodes. A *coding gain* of a network is the ratio of a coding capacity to a routing capacity. For directed multicast[2] and directed multiple unicast[3] networks, Sanders, Egner, and Tolhuizen [10] and Li and Li [8] respectively showed that the coding gain can be arbitrarily large.

An important problem is to determine how many nodes in a network are required to perform coding in order for the network to achieve its coding capacity (or to achieve a coding rate arbitrarily close to its capacity if the capacity is not actually achievable). A network node is said to be a *coding node* if at least one of its out-edges has a non-routing edge function. A similar problem is to determine the number of coding nodes needed to assure the network has a solution (i.e. a $(k_1, \ldots, k_m, n)$ fractional solution with $k_1 = \cdots = k_m = n = 1$). The number of required coding nodes in both problems can in general range anywhere from zero up to the total number of nodes in the network.

For the special case of multicast networks, the problem of finding a minimal set of coding nodes to solve a network has been examined previously in [2], [6], [7], [11]; the results of which are summarized as follows. Langberg, Sprintson, and Bruck [7] determined upper bounds on the minimum number of coding nodes required for a solution. Their bounds are given as functions of the number of messages and the number of receivers. Tavory, Feder, and Ron [11] showed that with two source messages, the minimum number of coding nodes required for a solution is independent of the total number of nodes in the network, while Fragouli and Soljanin [6] showed this minimum to be upper bounded by the number of receivers. Bhattad, Ratnakar, Koetter, and Narayanan [2] gave a method for finding solutions with reduced numbers of coding nodes, but their method may not find the minimum possible number of coding nodes. Wu, Jain, and Kung [12] demonstrated that only certain network edges require coding functions. This fact indirectly influences the number of coding nodes required, but does not immediately give an algorithm for finding a minimum node set.

We study here a related (and more general) problem, namely how network coding capacities can vary as functions of the number of allowable coding nodes. Our main re-

[2]A *multicast* network is a network with a single source and with every receiver demanding all of the source messages.

[3]A *multiple unicast* network is a network where each message is generated by exactly one source node and is demanded by exactly one receiver node.

sult, given in Theorem 3.3.2, shows that the capacities of networks, as functions of the number of allowable coding nodes, can be almost anything. That is, the class of directed acyclic networks can witness arbitrary amounts of coding gain by using arbitrarily-sized node subsets for coding.

3.2 Coding Capacities

Various coding capacities can be defined in terms of the achievable rate region of a network. We study two such quantities, presenting their definitions and determining their values for an example network given in Figure 3.1. This network is used to establish Theorem 3.3.2. Li and Li [8] presented a variation of this network and found the routing and coding capacities for the case when $k_i = k$ for all i.

For any $(k_1, \ldots, k_m, n)$ fractional solution, we call the scalar value

$$\frac{1}{m}\left(\frac{k_1}{n} + \cdots + \frac{k_m}{n}\right)$$

an *achievable average rate* of the network. We define the *average coding capacity* of a network to be the supremum of all achievable average rates, namely

$$\mathcal{C}^{average} = \sup\left\{\frac{1}{m}\sum_{i=1}^{m} r_i : (r_1, \ldots, r_m) \in S\right\}.$$

Similarly, for any $(k_1, \ldots, k_m, n)$ fractional solution, we call the scalar quantity

$$\min\left\{\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right\}$$

an *achievable uniform rate* of the network. We define the *uniform coding capacity* of a network to be the supremum of all achievable uniform rates, namely

$$\mathcal{C}^{uniform} = \sup\left\{\min_{1\le i\le m} r_i : (r_1, \ldots, r_m) \in S\right\}.$$

Note that if $r \in S$ and if $r' \in \mathbb{Q}^{m+}$ is component-wise less than or equal to r, then $r' \in S$. In particular, if

$$(r_1, \ldots, r_m) \in S$$

and

$$r_i = \min_{1\le j\le m} r_j$$

then

$$(r_i, r_i, \ldots, r_i) \in S$$

which implies

$$\mathcal{C}^{uniform} = \sup \{r_i : (r_1, \ldots, r_m) \in S, \;\; r_1 = \cdots = r_m\}.$$

In other words, all messages can be restricted to having the same dimension

$$k_1 = \cdots = k_m$$

when considering $\mathcal{C}^{uniform}$.

Also, note that

$$\mathcal{C}^{average} \geq \mathcal{C}^{uniform}$$

and that quantities $\mathcal{C}^{average}$ and $\mathcal{C}^{uniform}$ are attained by points on the boundary of the closure $\bar{S}$ of S. If a network's edge functions are restricted to purely routing functions, then $\mathcal{C}^{average}$ and $\mathcal{C}^{uniform}$ will be referred to as the *average routing capacity* and *uniform routing capacity*, and will be denoted $\mathcal{C}_0^{average}$ and $\mathcal{C}_0^{uniform}$, respectively.

Example 3.2.1. In this example, we consider the network in Figure 3.1. Note that for each $j = 1, \ldots, q$, every path from source node n_j to receiver node n_{q+2+j} contains the edge $e_{j,q+1}$. Thus, we must have $k_j \leq n$ for all j, and therefore

$$k_1 + \cdots + k_q \leq qn,$$

so $\mathcal{C}^{average} \leq 1$.

Furthermore, we can obtain a $(k_1, \ldots, k_q, n)$ fractional coding solution with

$$k_1 = \cdots = k_q = n = 1$$

using routing at all nodes except n_{q+1}, which transmits the mod $|\mathcal{A}|$ sum of its inputs on one of its out-edges and nothing on its other $p - 1$ out-edges. This solution implies that

$$\mathcal{C}^{average} \geq 1.$$

Thus, we have $\mathcal{C}^{average} = 1$.

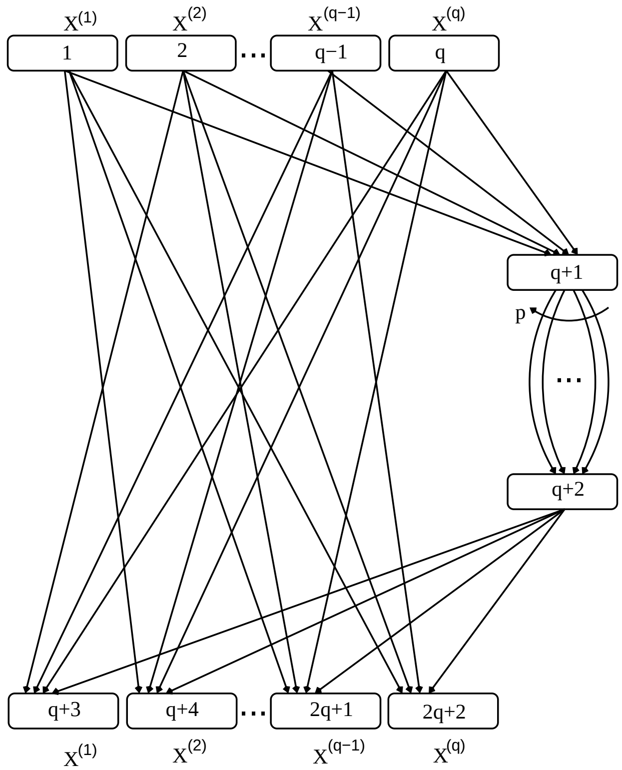

Figure 3.1: The network $\mathcal{N}(p,q)$, with $p \leq q$ and $p,q \in \mathbb{Z}^+$. Nodes $n_1,\dots,n_q$ are the sources, with node n_i providing message $X^{(i)}$, for $1 \leq i \leq q$. Nodes $n_{q+3},\dots,n_{2q+2}$ are the receivers, with node n_i demanding message $X^{(i-q-2)}$, for $q+3 \leq i \leq 2q+2$. Every source has one out-edge going to node n_{q+1} and every receiver has one in-edge coming from node n_{q+2}. Also, every source n_i has an out-edge going to receiver n_{q+2+j}, for all $j \neq i$. There are p parallel edges from node n_{q+1} to node n_{q+2}.

Clearly,

$$\mathcal{C}^{uniform} \leq \mathcal{C}^{average} = 1.$$

The presented $(k_1,\dots,k_q,n)$ fractional coding solution uses

$$k_1 = \cdots = k_q,$$

so

$$\mathcal{C}^{uniform} \geq 1.$$

Thus,

$$\mathcal{C}^{uniform} = 1.$$

When only routing is allowed, all of the messages must pass through the p edges from node n_{q+1} to n_{q+2}. Thus, we must have

$$k_1 + \cdots + k_q \leq pn,$$

or equivalently,

$$\frac{k_1 + \cdots + k_q}{qn} \leq \frac{p}{q}.$$

This implies

$$\mathcal{C}_0^{average} \leq \frac{p}{q}.$$

A $(k_1, \ldots, k_q, n)$ fractional routing solution consists of taking

$$k_1 = \cdots = k_q = p$$

and $n = q$ and sending each message $X^{(j)}$ along the corresponding edge $e_{j,q+1}$, sending all

$$k_1 + \cdots + k_q = qp$$

message components from node n_{q+1} to n_{q+2} in an arbitrary fashion, and then sending each message $X^{(j)}$ from node n_{q+2} to the corresponding receiver node n_{q+2+j}. Hence,

$$\mathcal{C}_0^{uniform} \geq \frac{p}{q}$$

and therefore

$$\frac{p}{q} \leq \mathcal{C}_0^{uniform} \leq \mathcal{C}_0^{average} \leq \frac{p}{q}.$$

Thus,

$$\mathcal{C}_0^{uniform} = \mathcal{C}_0^{average} = \frac{p}{q}.$$

Various properties of network routing and coding capacities relating to their relative values, linearity, alphabet size, achievability, and computability have previously been studied [1], [3]–[5], [9]. However, it is not presently known whether or not there exist algorithms that can compute the coding capacity (uniform or average) of an arbitrary network. In fact, computing the exact coding capacity of even relatively simple networks can be a seemingly non-trivial task. At present, very few exact coding capacities have been rigorously derived in the literature.

3.3 Node-Limited Coding Capacities

For each non-negative integer i, a $(k_1, \ldots, k_m, n)$ *fractional i-node coding solution* for a network is a $(k_1, \ldots, k_m, n)$ *fractional coding solution* with at most i coding nodes (i.e. having output edges with non-routing edge functions).[4] For each i, we denote by $\mathcal{C}_i^{average}$ and $\mathcal{C}_i^{uniform}$ the average and uniform coding capacities, respectively, when solutions are restricted to those having at most i coding nodes. We make the convention that, for all $i > |V|$,

$$\mathcal{C}_i^{average} = \mathcal{C}_{|V|}^{average}$$

and

$$\mathcal{C}_i^{uniform} = \mathcal{C}_{|V|}^{uniform}.$$

We call $\mathcal{C}_i^{average}$ and $\mathcal{C}_i^{uniform}$ the *node-limited average capacity function* and *node-limited uniform capacity function*, respectively. Clearly, the minimum number of coding nodes needed to obtain the average or uniform network capacity is the smallest i such that

$$\mathcal{C}_i^{average} = \mathcal{C}^{average}$$

or

$$\mathcal{C}_i^{uniform} = \mathcal{C}^{uniform},$$

respectively. Also, the quantities $\mathcal{C}_{|V|}^{uniform}$ and $\mathcal{C}_{|V|}^{average}$ are respectively the uniform and average coding capacities.

Example 3.3.1. For the network in Figure 3.1, since $\mathcal{C}^{average}$ and $\mathcal{C}^{uniform}$ are both achieved using only a single coding node (as shown in Example 3.2.1), the node-limited capacities are

$$\mathcal{C}_i^{average} = \mathcal{C}_i^{uniform} = \begin{cases} p/q & \text{for } i = 0 \\ 1 & \text{for } i \geq 1. \end{cases} \tag{3.1}$$

A function $f : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is said to be *eventually constant* if there exists an i such that

$$f(i + j) = f(i)$$

[4] Arbitrary decoding is allowed at receiver nodes and receiver nodes only contribute to the total number of coding nodes in a network if they have out-edges performing coding.

for all $j \in \mathbb{N}$. Thus, the node-limited uniform and average capacity functions are eventually constant. A network's node-limited capacity function is also always non-negative. For a given number of coding nodes, if a network's node-limited capacity is achievable, then it must be rational, and cannot decrease if more nodes are allowed to perform coding (since one can choose not to use extra nodes for coding). By examining the admissible forms of $\mathcal{C}_i^{average}$ and $\mathcal{C}_i^{uniform}$ we gain insight into the possible capacity benefits of performing network coding at a limited number of nodes.

Theorem 3.3.2, whose proof appears after Lemma 3.3.4, demonstrates that node-limited capacities of networks can vary more-or-less arbitrarily as functions of the number of allowable coding nodes. Thus, there cannot exist any useful general upper or lower bounds on the node-limited capacity of an arbitrary network (bounds might exist as functions of the properties of specific networks, however).

Theorem 3.3.2. *Every monotonically non-decreasing, eventually constant function $f : \mathbb{N} \cup \{0\} \to \mathbb{Q}^+$ is the node-limited average and uniform capacity function of some directed acyclic network.*

Two lemmas are now stated (the proofs are simple and therefore omitted) and are then used to prove Theorem 3.3.2.

Lemma 3.3.3. *Let $\mathcal{N}$ be a network with node-limited uniform and average coding capacities $\mathcal{C}_i^{uniform}$ and $\mathcal{C}_i^{average}$, respectively, and let p be a positive integer. If every message is replaced at its source node by p new independent messages and every receiver has each message demand replaced by a demand for all of the p new corresponding messages, then the node-limited uniform and average coding capacity functions of the resulting network $\mathcal{N}'$ are $(1/p)\mathcal{C}_i^{uniform}$ and $(1/p)\mathcal{C}_i^{average}$, respectively.*

Lemma 3.3.4. *Let $\mathcal{N}$ be a network with node-limited uniform and average coding capacities $\mathcal{C}_i^{uniform}$ and $\mathcal{C}_i^{average}$, respectively, and let q be a positive integer. If every directed edge is replaced by q new parallel directed edges in the same orientation, then the node-limited uniform and average coding capacity functions of the resulting network $\mathcal{N}'$ are $q\mathcal{C}_i^{uniform}$ and $q\mathcal{C}_i^{average}$, respectively.*

Proof of Theorem 3.3.2.

Suppose $f : \mathbb{N} \cup \{0\} \to \mathbb{Q}^+$ is given by

$$f(i) = \begin{cases} p_i/q_i & \text{for } 0 \leq i < s \\ p_s/q_s & \text{for } i \geq s \end{cases}$$

where

$$p_0, \ldots, p_s, q_0, \ldots, q_s$$

are positive integers such that

$$p_0/q_0 \leq p_1/q_1 \leq \cdots \leq p_s/q_s.$$

Define the positive integers

$$b = p_s \cdot \mathrm{lcm}\{q_i : 0 \leq i < s\} = \mathrm{lcm}\{p_s q_i : 0 \leq i < s\} \in \mathbb{N}$$
$$a_i = \frac{p_i/q_i}{p_s/q_s} \cdot b = \frac{p_i q_s}{p_s q_i} \cdot b \in \mathbb{N}$$

and construct a network $\mathcal{N}$ as shown in Figure 3.2, which has $m = b$ source messages and uses the networks

$$\mathcal{N}(a_0, b), \ldots, \mathcal{N}(a_{s-1}, b)$$

as building blocks (note that $a_i/b \leq 1$ for all i).

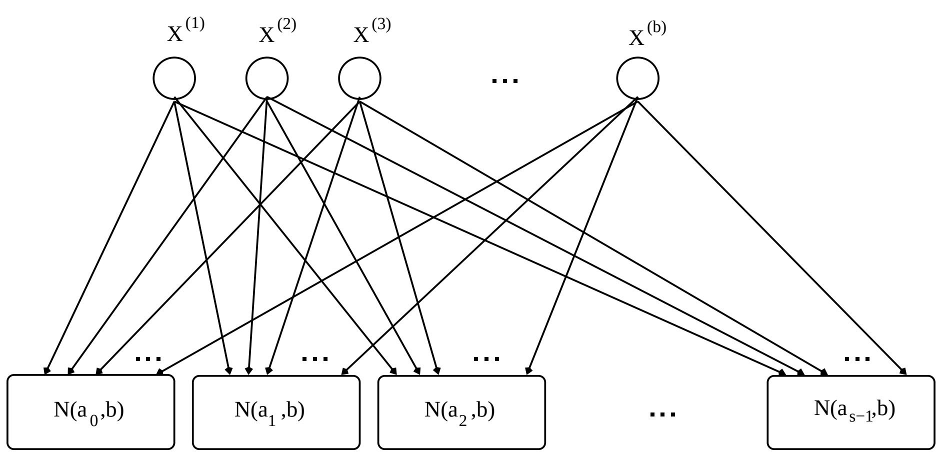

Figure 3.2: The network $\mathcal{N}$ has b source nodes, each emitting one message. Each source node has an out-edge to each sub-block $\mathcal{N}(a_0, b), \ldots, \mathcal{N}(a_{s-1}, b)$. Specifically, in each sub-block $\mathcal{N}(a_i, b)$, the previous source messages are removed, however each previous source node is connected by an in-edge from the unique corresponding source node in $\mathcal{N}$. Each sub-block $\mathcal{N}(a_i, b)$ has routing capacity $a_i/b = (p_i/q_i)/(p_s/q_s)$.

Let $\mathcal{C}_i^{uniform}$ and $\mathcal{C}_i^{average}$ denote the uniform and average node-limited capacity functions of network $\mathcal{N}$. Also, for $j = 0, \ldots, s-1$, let $C_{j,i}^{uniform}$ and $C_{j,i}^{average}$ denote the uniform and average node-limited capacity functions of the sub-block $\mathcal{N}(a_j, b)$. There are exactly $2s$ nodes in $\mathcal{N}$ that have more than one in-edge and at least one out-edge, and which are therefore potential coding nodes (i.e. two potential coding nodes per sub-block). However, for each sub-block, any coding performed at the lower potential coding node can be directly incorporated into the upper potential coding node.

For each $i = 0, \ldots, s-1$, in order to obtain a $(k_1, \ldots, k_m, n)$ fractional i-node coding solution, the quantity

$$\frac{k_1 + \cdots + k_m}{mn}$$

must be at most

$$\min_j \frac{a_j}{b} = \min_j \frac{p_j/q_j}{p_s/q_s}$$

where the minimization is taken over all j for which sub-block $\mathcal{N}(a_j, b)$ has no coding nodes (as seen from (3.1)). That is, we must have

$$\frac{k_1 + \cdots + k_m}{mn} \leq \frac{p_i/q_i}{p_s/q_s}.$$

Therefore, the node-limited average and uniform coding capacities of $\mathcal{N}$ using i coding nodes are at most the respective routing capacities of sub-block $\mathcal{N}(a_i, b)$ of $\mathcal{N}$, namely

$$\mathcal{C}_i^{uniform} \leq C_{i,0}^{uniform} = a_i/b = \frac{p_i/q_i}{p_s/q_s}$$
$$\mathcal{C}_i^{average} \leq C_{i,0}^{average} = a_i/b = \frac{p_i/q_i}{p_s/q_s}.$$

These upper bounds are achievable by using coding at the one useful possible coding node in each of the sub-blocks

$$\mathcal{N}(a_0, b), \ldots, \mathcal{N}(a_{i-1}, b)$$

and using routing elsewhere. By taking

$$d = \text{lcm}(a_i, \ldots, a_{s-1})$$
$$k_1 = \cdots = k_m = d$$
$$n = bd/a_i$$

we can obtain a $(k_1, \ldots, k_m, n)$ fractional i-node coding solution with coding nodes in sub-blocks

$$\mathcal{N}(a_0, b), \ldots, \mathcal{N}(a_{i-1}, b)$$

and only routing edge-functions in sub-blocks

$$\mathcal{N}(a_i, b), \ldots, \mathcal{N}(a_{s-1}, b).$$

With such a solution, the coding capacity

$$C_{j,1}^{uniform} = C_{j,1}^{average} = 1$$

is achieved in each sub-block

$$\mathcal{N}(a_0, b), \ldots, \mathcal{N}(a_{i-1}, b),$$

and the (unchanging) routing capacity

$$C_{i,0}^{uniform} = C_{i,0}^{average}$$

is achieved in each sub-block

$$\mathcal{N}(a_i, b), \ldots, \mathcal{N}(a_{s-1}, b).$$

Thus, network $\mathcal{N}$ has node-limited average and uniform capacity functions given by

$$\mathcal{C}_i^{average} = \mathcal{C}_i^{uniform} = \begin{cases} (p_i/q_i)/(p_s/q_s) & \text{for } 0 \le i < s \\ 1 & \text{for } i \ge s. \end{cases}$$

By Lemma 3.3.3 and Lemma 3.3.4, if we replace each message of $\mathcal{N}$ by q_s new independent messages and change the receiver demands accordingly, and if we replace each directed edge of $\mathcal{N}$ by p_s parallel edges in the same orientation, then the resulting network $\hat{\mathcal{N}}$ will have node-limited average and uniform capacity functions given by

$$\hat{\mathcal{C}}_i^{average} = \hat{\mathcal{C}}_i^{uniform} = (p_s/q_s)\mathcal{C}_i^{uniform} = f(i).$$

■

We note that a simpler network could have been used in the proof of Theorem 3.3.2 if only the case of $\mathcal{C}_i^{uniform}$ were considered. Namely, we could have used only $\max_{O \le i < s} q_i p_s$ source nodes and then connected edges from source nodes to sub-blocks $\mathcal{N}(p_i q_s, q_i p_s)$ as needed.

One consequence of Theorem 3.3.2 is that large coding gains can be suddenly obtained after an arbitrary number of nodes has been used for coding. For example, for any integer $i \ge 0$ and for any real number $t > 0$, there exists a network such that

$$\begin{aligned}
&\mathcal{C}_0^{uniform} = \mathcal{C}_1^{uniform} = \cdots = \mathcal{C}_i^{uniform} \\
&\mathcal{C}_0^{average} = \mathcal{C}_1^{average} = \cdots = \mathcal{C}_i^{average} \\
&\mathcal{C}_{i+1}^{uniform} - \mathcal{C}_i^{uniform} > t \\
&\mathcal{C}_{i+1}^{average} - \mathcal{C}_i^{average} > t.
\end{aligned}$$

In Theorem 3.3.2 the existence of networks that achieve prescribed rational-valued node-limited capacity functions was established. It is known in general that not all networks necessarily achieve their capacities [5]. It is presently unknown, however, whether a network coding capacity could be irrational.[5] Thus, we are not presently able to extend Theorem 3.3.2 to real-valued functions. Nevertheless, Theorem 3.3.2 does immediately imply the following asymptotic achievability result for real-valued functions.

Corollary 3.3.5. *Every monotonically non-decreasing, eventually constant function* f : $\mathbb{N} \cup \{0\} \to \mathbb{R}^+$ *is the limit of the node-limited uniform and average capacity function of some sequence of directed acyclic networks.*

3.4 Acknowledgment

The text of this chapter, in full, is a reprint of the material as it appears in Jillian Cannons and Kenneth Zeger, “Network Coding Capacity With a Constrained Number of Coding Nodes,” *IEEE Transactions on Information Theory*, vol. 54, no. 3, March 2008.

[5] It would be interesting to understand whether, for example, a node-limited capacity function of a network could take on some rational and some irrational values, and perhaps achieve some values and not achieve other values. We leave this as an open question.

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow", *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204 – 1216, Jul. 2000.

[2] K. Bhattad, N. Ratnakar, R. Koetter, and K. R. Narayanan, "Minimal network coding for multicast", in *Proc. 2005 IEEE Int. Symp. Information Theory (ISIT)*, Adelaide, Australia, Sep. 2005.

[3] J. Cannons, R. Dougherty, C. Freiling, and K. Zeger, "Network routing capacity", *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 777 – 788, Mar. 2006.

[4] R. Dougherty, C. Freiling, and K. Zeger, "Insufficiency of linear coding in network information flow", *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2745 – 2759, Aug. 2005.

[5] R. Dougherty, C. Freiling, and K. Zeger, "Unachievability of network coding capacity", *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2365 – 2372, June 2006, Joint issue with *IEEE/ACM Trans. Netw*.

[6] C. Fragouli and E. Soljanin, "Information flow decomposition for network coding", *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 829 – 848, Mar. 2006.

[7] M. Langberg, A. Sprintson, and J. Bruck, "The encoding complexity of network coding", *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2386 – 2397, Jun. 2006, Joint issue with *IEEE/ACM Trans. Netw*.

[8] Z. Li and B. Li, "Network coding: The case of multiple unicast sessions", in *Proc. 42nd Ann. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2004.

[9] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding", *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371 – 381, Feb. 2003.

[10] P. Sanders, S. Egner, and L. Tolhuizen, "Polynomial time algorithms for network information flow", in *Proc. 15th Ann. ACM Symp. Parallelism in Algorithms and Architectures (SPAA)*, San Diego, CA, Jun. 2003, pp. 286 – 294.

[11] A. Tavory, M. Feder, and D. Ron, "Bounds on linear codes for network multicast", *Proc. Electronic Colloquium on Computational Complexity (ECCC)*, pp. 1 – 28, 2003.

[12] Y. Wu, K. Jain, and S.-Y. Kung, "A unification of network coding and tree packing (routing) theorems", *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2398 – 2409, Jun. 2006, Joint issue with *IEEE/ACM Trans. Netw*.

[13] R. W. Yeung, *A First Course in Information Theory*, Amsterdam, The Netherlands: Kluwer, 2002.

Chapter 4

An Algorithm for Wireless Relay Placement

Abstract

An algorithm is given for placing relays at spatial positions to improve the reliability of communicated data in a sensor network. The network consists of many power-limited sensors, a small set of relays, and a receiver. The receiver receives a signal directly from each sensor and also indirectly from one relay per sensor. The relays rebroadcast the transmissions in order to achieve diversity at the receiver. Both amplify-and-forward and decode-and-forward relay networks are considered. Channels are modeled with Rayleigh fading, path loss, and additive white Gaussian noise. Performance analysis and numerical results are given.

4.1 Introduction

Wireless sensor networks typically consist of a large number of small, power-limited sensors distributed over a planar geographic area. In some scenarios, the sensors collect information which is transmitted to a single receiver for further analysis. A small number of radio relays with additional processing and communications capabilities can be strategically placed to help improve system performance. Two important problems we consider here are to position the relays and to determine, for each sensor, which relay should rebroadcast its signal.

Previous studies of relay placement have considered various optimization criteria and communication models. Some have focused on the coverage of the network (e.g., Balam and Gibson [2]; Chen, Wang, and Liang [4]; Cortés, Martiínez, Karataş, and Bullo [7]; Koutsopoulos, Toumpis, and Tassiulas [13]; Liu and Mohapatra [14]; Mao and Wu [15]; Suomela [22]; Tan, Lozano, Xi, and Sheng [23]). In [13] communication errors are modeled by a fixed probability of error without incorporating physical considerations; otherwise, communications are assumed to be error-free. Such studies often directly use the source coding technique known as the Lloyd algorithm (e.g., see [9]), which is sub-optimal for relay placement. Two other optimization criteria are network lifetime and energy usage, with energy modeled as an increasing function of distance and with error-free communications (e.g., Ergen and Varaiya [8]; Hou, Shi, Sherali, and Midkiff [11]; Iranli, Maleki, and Pedram [12]; Pan, Cai, Hou, Shi, and Shen [17]). Models incorporating fading and/or path loss have been used for criteria such as error probability, outage probability, and throughput, typically with simplifications such as single-sensor or single-relay networks (e.g., Cho and Yang [5]; So and Liang [21]; Sadek, Han, and Liu [20]). The majority of the above approaches do not include diversity. Those that do often do not focus on optimal relay location and use restricted networks with only a single source and/or a single relay (e.g., Ong and Motani [16]; Chen and Laneman [3]). These previous studies offer valuable insight; however, the communication and/or network models used are typically simplified.

In this work, we attempt to position the relays and determine which relay should rebroadcast each sensor's transmissions in order to minimize the average probability of error. We use a more elaborate communications model which includes path loss, fading, additive white Gaussian noise, and diversity. We use a network model in which all relays either use amplify-and-forward or decode-and-forward communications. Each sensor in the network transmits information to the receiver both directly and through a single-hop relay path. The receiver uses the two received signals to achieve diversity. Sensors identify themselves in transmissions and relays know for which sensors they are responsible. We assume TDMA communications by sensors and relays so that there is (ideally) no transmission interference.

We present an algorithm that determines relay placement and assigns each sensor to a relay. We refer to this algorithm as the *relay placement algorithm*. The algorithm has some similarity to the Lloyd algorithm. We describe geometrically, with respect to fixed relay positions, the sets of locations in the plane in which sensors are (optimally) assigned to the same relay, and give performance results based on these analyses and using numerical computations.

In Section 4.2, we specify communications models and determine error probabilities. In Section 4.3, we present our relay placement algorithm. In Section 4.4, we give analytic descriptions of optimal sensor regions (with respect to fixed relay positions). In Section 4.5, we present numerical results. In Section 4.6, we summarize our work and provide ideas for future consideration.

4.2 Communications Model and Performance Measure

4.2.1 Signal, Channel, and Receiver Models

In a sensor network, we refer to sensors, relays, and the receiver as *nodes*. We assume that transmission of $b_i \in \{-1, 1\}$ by node i uses the binary phase shift keyed (BPSK) signal $s_i(t)$, and we denote the transmission energy per bit by E_i. In particular, we assume all sensor nodes transmit at the same energy per bit, denoted by E_{Tx}. The communications channel model includes path loss, additive white Gaussian noise (AWGN), and fading. Let $L_{i,j}$ denote the far field path loss between two nodes i and j that are separated by a distance $d_{i,j}$ (in meters). We consider the free-space law model (e.g., see [19, pp. 70 – 73]) for which[1]

$$L_{i,j} = \frac{F_2}{d_{i,j}^2} \tag{4.1}$$

where:

$F_2 = \frac{\lambda^2}{16\pi^2}$ (in meters^2)

[1]Much of the material of this paper can be generalized by replacing the path loss exponent 2 by any positive, even integer, and F_2 by a corresponding constant.

$\lambda = c/f_0$ is the wavelength of the carrier wave (in meters)

$c = 3 \cdot 10^8$ is the speed of light (in meters/second)

f_0 is the frequency of the carrier wave (in Hz).

The formula in (4.1) is impractical in the near field, since $L_{i,j} \to \infty$ as $d_{i,j} \to 0$. Comaniciu and Poor [6] addressed this issue by not allowing transmissions at distances less than λ. Ong and Motani [16] allow near field transmissions by proposing a modified model with path loss

$$L_{i,j} = \frac{F_2}{(1 + d_{i,j})^2}. \tag{4.2}$$

We assume additive white Gaussian noise $n_j(t)$ at the receiving antenna of node j. The noise has one-sided power spectral density N_0 (in W/Hz). Assume the channel fading (excluding path loss) between nodes i and j is a random variable $h_{i,j}$ with Rayleigh density

$$p_{h_{i,j}}(h) = (h/\sigma^2)e^{-h^2/(2\sigma^2)} \quad (h \geq 0). \tag{4.3}$$

We also consider AWGN channels (which is equivalent to assuming $h_{i,j} = 1$ for all i, j).

Let the signal received after transmission from node i to node j be denoted by $r_{i,j}(t)$. Combining the signal and channel models, we have $r_{i,j}(t) = \sqrt{L_{i,j}}\ h_{i,j} s_i(t) + n_j(t)$. The received energy per bit without fading is $E_j = E_i L_{i,j}$. We assume demodulation at a receiving node is performed by applying a matched filter to obtain the test statistic. Diversity is achieved at the receiver by making a decision based on a test statistic that combines the two received versions (i.e., direct and relayed) of the transmission from a given sensor. We assume the receiver uses selection combining, in which only the better of the two incoming signals (determined by a measurable quantity such as the received signal-to-noise-ratio (SNR)) is used to detect the transmitted bit.

4.2.2 Path Probability of Error

For each sensor, we determine the probability of error along the direct path from the sensor to the receiver and along single-hop[2] relay paths, for both amplify-and-forward and decode-and-forward protocols. Let $\mathbf{x} \in \mathbb{R}^2$ denote a transmitter position and let Rx

[2]Computing the probabilities of error for the more general case of multi-hop relay paths is straightforward.

denote the receiver. We consider transmission paths of the forms $(\mathbf{x}, \mathrm{Rx})$, $(\mathbf{x}, i)$, (i, Rx), and $(\mathbf{x}, i, \mathrm{Rx})$, where i denotes a relay index. For each such path q, let:

$$\mathrm{SNR}_H^q = \text{end-to-end SNR, conditioned on the fades} \tag{4.4}$$

$$P_{e|H}^q = \text{end-to-end error probability, conditioned on the fades} \tag{4.5}$$

$$\mathrm{SNR}^q = \text{end-to-end SNR} \tag{4.6}$$

$$P_e^q = \text{end-to-end error probability.} \tag{4.7}$$

For AWGN channels, we take SNR^q and P_e^q to be the SNR and error probability when the signal is degraded only by path loss and receiver antenna noise. For fading channels, we take SNR^q and P_e^q to also be averaged over the fades. Note that the signal-to-noise ratios only apply to direct paths and paths using amplify-and-forward relays. Finally, denote the Gaussian error function by $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy$.

Direct Path (i.e., unrelayed)

For Rayleigh fading, we have (e.g., see [18, pp. 817 – 818])

$$\mathrm{SNR}^{(\mathbf{x},\mathrm{Rx})} = \frac{4\sigma^2 E_{\mathrm{Tx}} L_{\mathbf{x},\mathrm{Rx}}}{N_0}; \qquad \mathrm{SNR}^{(\mathbf{x},i)} = \frac{4\sigma^2 E_{\mathrm{Tx}} L_{\mathbf{x},i}}{N_0}; \qquad \mathrm{SNR}^{(i,\mathrm{Rx})} = \frac{4\sigma^2 E_i L_{i,\mathrm{Rx}}}{N_0} \tag{4.8}$$

$$P_e^{(\mathbf{x},\mathrm{Rx})} = \frac{1}{2}\left(1 - \left(1 + \frac{2}{\mathrm{SNR}^{(\mathbf{x},\mathrm{Rx})}}\right)^{-1/2}\right). \tag{4.9}$$

For AWGN channels, we have (e.g., see [18, pp. 255 – 256])

$$\mathrm{SNR}^{(\mathbf{x},\mathrm{Rx})} = \frac{2E_{\mathrm{Tx}} L_{\mathbf{x},\mathrm{Rx}}}{N_0}; \qquad \mathrm{SNR}^{(\mathbf{x},i)} = \frac{2E_{\mathrm{Tx}} L_{\mathbf{x},i}}{N_0}; \qquad \mathrm{SNR}^{(i,\mathrm{Rx})} = \frac{2E_i L_{i,\mathrm{Rx}}}{N_0} \tag{4.10}$$

$$P_e^{(\mathbf{x},\mathrm{Rx})} = Q\left(\sqrt{\mathrm{SNR}^{(\mathbf{x},\mathrm{Rx})}}\right). \tag{4.11}$$

Note that analogous formulas to those in (4.9) and (4.11) can be given for $P_e^{(\mathbf{x},i)}$ and $P_e^{(i,\mathrm{Rx})}$.

Relay Path with Amplify-and-Forward

For amplify-and-forward relays,[3] the system is linear. Denote the gain by G. Conditioning on the fading values, we have (e.g., see [10])

$$\mathrm{SNR}_H^{(\mathbf{x},i,\mathrm{Rx})} = \frac{h_{\mathbf{x},i}^2 h_{i,\mathrm{Rx}}^2 E_{\mathrm{Tx}}/N_0}{B_i h_{i,\mathrm{Rx}}^2 + D_i} \tag{4.12}$$

$$P_{e|H}^{(\mathbf{x},i,\mathrm{Rx})} = Q\left(\sqrt{\mathrm{SNR}_h^{(\mathbf{x},i,\mathrm{Rx})}}\right) \tag{4.13}$$

$$\text{where } B_i = \frac{1}{2L_{\mathbf{x},i}}; \qquad D_i = \frac{1}{2G^2 L_{\mathbf{x},i} L_{i,\mathrm{Rx}}}. \tag{4.14}$$

Then, the end-to-end probability of error, averaged over the fades, is

$$\begin{aligned}
P_e^{(\mathbf{x},i,\mathrm{Rx})} &= \int_0^\infty \int_0^\infty P_{e|H}^{(\mathbf{x},i,\mathrm{Rx})} p_H\left(h_{\mathbf{x},i}\right) p_H\left(h_{i,\mathrm{Rx}}\right) dh_{\mathbf{x},i}\, dh_{i,\mathrm{Rx}} \\
&= \int_0^\infty \int_0^\infty Q\left(\sqrt{\frac{h_{\mathbf{x},i}^2 h_{i,\mathrm{Rx}}^2 E_{\mathrm{Tx}}/N_0}{B_i h_{i,\mathrm{Rx}}^2 + D_i}}\right) \frac{h_{\mathbf{x},i}}{\sigma^2} \cdot \exp\left\{-\frac{h_{\mathbf{x},i}^2}{2\sigma^2}\right\} \frac{h_{i,\mathrm{Rx}}}{\sigma^2} \\
&\qquad\qquad \cdot \exp\left\{-\frac{h_{i,\mathrm{Rx}}^2}{2\sigma^2}\right\} dh_{\mathbf{x},i}\, dh_{i,\mathrm{Rx}} \quad \text{[from (4.13), (4.12), (4.3)]} \\
&= \frac{1}{2} - \frac{D_i N_0/E_{\mathrm{Tx}}}{4\sigma\left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)^{3/2}} \\
&\qquad \cdot \int_0^\infty \sqrt{\frac{t}{t+1}} \cdot \exp\left\{-t\left(\frac{D_i N_0/E_{\mathrm{Tx}}}{2\sigma^2\left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)}\right)\right\} dt \\
&= \frac{1}{2} - \frac{D_i \sqrt{\pi} N_0/E_{\mathrm{Tx}}}{8\sigma\left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)^{3/2}} \cdot U\left(\frac{3}{2}, 2, \frac{D_i N_0/E_{\mathrm{Tx}}}{2\sigma^2\left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)}\right)
\end{aligned} \tag{4.15}$$

where $U(a,b,z)$ denotes the confluent hypergeometric function of the second kind [1, p. 505] (also known as Kummer's function of the second kind), i.e.,

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1}\, dt.$$

For AWGN channels, we have

$$\mathrm{SNR}^{(\mathbf{x},i,\mathrm{Rx})} = \frac{E_{\mathrm{Tx}}/N_0}{B_i + D_i} \qquad \text{[from (4.12)]} \tag{4.16}$$

$$P_e^{(\mathbf{x},i,\mathrm{Rx})} = Q\left(\sqrt{\mathrm{SNR}^{(\mathbf{x},i,\mathrm{Rx})}}\right). \tag{4.17}$$

[3]By *amplify-and-forward relays* we specifically mean that a received signal is multiplied by a constant gain factor and then transmitted.

Relay Path with Decode-and-Forward

For decode-and-forward relays,[4] the signal at the receiver is not a linear function of the transmitted signal (i.e., the system is not linear), as the relay makes a hard decision based on its incoming data. A decoding error occurs at the receiver if and only if exactly one decoding error is made along the relay path. Thus, for Rayleigh fading, we obtain (e.g., see [10])

$$P_e^{(\mathbf{x},i,\mathrm{Rx})} = \frac{1}{4}\left(1 - \left(1 + \frac{2}{\mathrm{SNR}^{(\mathbf{x},i)}}\right)^{-1/2}\right)\left(1 + \left(1 + \frac{2}{\mathrm{SNR}^{(i,\mathrm{Rx})}}\right)^{-1/2}\right) + \frac{1}{4}\left(1 - \left(1 + \frac{2}{\mathrm{SNR}^{(i,\mathrm{Rx})}}\right)^{-1/2}\right)\left(1 + \left(1 + \frac{2}{\mathrm{SNR}^{(\mathbf{x},i)}}\right)^{-1/2}\right). \quad [\text{from } (4.9)] \tag{4.18}$$

For AWGN channels, we have (e.g., see [10])

$$P_e^{(\mathbf{x},i,\mathrm{Rx})} = P_e^{(\mathbf{x},i)}\left(1 - P_e^{(i,\mathrm{Rx})}\right) + P_e^{(i,\mathrm{Rx})}\left(1 - P_e^{(\mathbf{x},i)}\right). \tag{4.19}$$

4.3 Path Selection and Relay Placement Algorithm

4.3.1 Definitions

We define a *sensor network with relays* to be a collection of sensors and relays in $\mathbb{R}^2$, together with a single receiver at the origin, where each sensor transmits to the receiver both directly and through some predesignated relay for the sensor, and the system performance is evaluated using the measure given below in (4.20). Specifically, let $\mathbf{x}_1, \ldots, \mathbf{x}_M \in \mathbb{R}^2$ be the sensor positions and let $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^2$ be the relay positions. Typically, $N \ll M$. Let $p : \mathbb{R}^2 \to \{1, \ldots, N\}$ be a *sensor-relay assignment*, where $p(\mathbf{x}) = i$ means that if a sensor were located at position $\mathbf{x}$, then it would be assigned to relay $\mathbf{y}_i$. Let $\mathcal{S}$ be a bounded subset of $\mathbb{R}^2$. Throughout this section and Section 4.4 we will consider sensor-relay assignments whose domains are restricted to $\mathcal{S}$ (since the number of sensors is finite).

[4] By *decode-and-forward relays* we specifically mean that a single symbol is demodulated and then re-modulated; no additional decoding is performed (e.g., of channel codes).

Let the *sensor-averaged probability of error* be given by

$$\frac{1}{M}\sum_{s=1}^{M} P_e^{(\mathbf{x}_s, p(\mathbf{x}_s), \mathrm{Rx})}. \tag{4.20}$$

Note that (4.20) depends on the relay locations through the sensor-relay assignment p. Finally, let $\langle\ ,\ \rangle$ denote the inner product operator.

4.3.2 Overview of the Proposed Algorithm

The proposed iterative algorithm attempts to minimize the sensor-averaged probability of error[5] over all choices of relay positions $\mathbf{y}_1, \ldots, \mathbf{y}_N$ and sensor-relay assignments p. The algorithm operates in two phases. First, the relay positions are fixed and the best sensor-relay assignment is determined; second, the sensor-relay assignment is fixed and the best relay positions are determined. An initial placement of the relays is made either randomly or using some heuristic. The two phases are repeated until the quantity in (4.20) has converged within some threshold.

4.3.3 Phase 1: Optimal Sensor-Relay Assignment

In the first phase, we assume the relay positions $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are fixed and choose an optimal[6] sensor-relay assignment p^*, in the sense of minimizing (4.20). This choice can be made using an exhaustive search in which all possible sensor-relay assignments are examined. A sensor-relay assignment induces a partition of $\mathcal{S}$ into subsets for which all sensors in any such subset are assigned to the same relay. For each relay $\mathbf{y}_i$, let σ_i be the set of all points $\mathbf{x} \in \mathcal{S}$ such that if a sensor were located at position $\mathbf{x}$, then the optimally assigned relay that rebroadcasts its transmissions would be $\mathbf{y}_i$, i.e., $\sigma_i = \{\mathbf{x} \in \mathcal{S} : p^*(\mathbf{x}) = i\}$. We call σ_i the i^{th} *optimal sensor region* (with respect to the fixed relay positions).

[5] Here we minimize (4.20); however, the algorithm can be adapted to minimize other performance measures.

[6] This choice may not be unique, but we select one such minimizing assignment here. Also, optimality of p^* here depends only on the values $p^*(\mathbf{x}_1), \ldots, p^*(\mathbf{x}_M)$.

4.3.4 Phase 2: Optimal Relay Placement

In the second phase, we assume the sensor-relay assignment is fixed and choose optimal[7] relay positions in the sense of minimizing (4.20). Numerical techniques can be used to determine such optimal relay positions. For the first three instances of phase 2 in the iterative algorithm, we used an efficient (but slightly sub-optimal) numerical approach that quantizes a bounded subset of $\mathbb{R}^2$ into gridpoints. For a given relay, the best gridpoint was selected as the new location for the relay. For subsequent instances of phase 2, the restriction of lying on a gridpoint was removed and a steepest descent technique was used to refine the relay locations.

4.4 Geometric Descriptions of Optimal Sensor Regions

We now geometrically describe each optimal sensor region by considering specific relay protocols and channel models. In particular, we examine amplify-and-forward and decode-and-forward relaying protocols in conjunction with either AWGN channels or Rayleigh fading channels. We define the *internal boundary* of any optimal sensor region σ_i to be the portion of the boundary of σ_i that does not lie on the boundary of $\mathcal{S}$. For amplify-and-forward and AWGN channels, we show that the internal boundary of each optimal sensor region consists only of circular arcs. For the other three combinations of relay protocol and channel type, we show that as the transmission energies of sensors and relays grow, the internal boundary of each optimal sensor region converges to finite combinations of circular arcs and/or line segments.

For each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$, let $\sigma_{i,j}$ be the set of all points $\mathbf{x} \in \mathcal{S}$ such that if a sensor were located at position $\mathbf{x}$, then its average probability of error using relay $\mathbf{y}_i$ would be smaller than that using relay $\mathbf{y}_j$, i.e.,

$$\sigma_{i,j} = \left\{ \mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})} \right\}. \tag{4.21}$$

Note that $\sigma_{i,j} = \mathcal{S} - \sigma_{j,i}$. Then, for the given set of relay positions, we have

$$\sigma_i = \bigcap_{\substack{j=1 \\ j \neq i}}^{N} \sigma_{i,j} \tag{4.22}$$

[7]This choice may not be unique, but we select one such set of positions here.

since $p^*(\mathbf{x}) = \underset{j\in\{1,\dots,N\}}{\operatorname{argmin}} P_e^{(\mathbf{x},j,\mathrm{Rx})}$. Furthermore, for a suitably chosen constant $C > 0$, in order to facilitate analysis, we modify (4.2) to[8]

$$L_{i,j} = \frac{F_2}{C + d_{i,j}^2}. \tag{4.23}$$

Amplify-and-Forward with AWGN Channels

Theorem 4.4.1. *Consider a sensor network with amplify-and-forward relays and AWGN channels. Then, the internal boundary of each optimal sensor region consists of circular arcs.*

Proof. For any distinct relays $\mathbf{y}_i$ and $\mathbf{y}_j$, let

$$K_i = \frac{1}{G^2 F_2 + C + \|\mathbf{y}_i\|^2}; \qquad \gamma_{i,j} = \frac{K_i}{K_i - K_j}. \tag{4.24}$$

Note that for fixed gain G, $K_i \neq K_j$ since we assume $\mathbf{y}_i \neq \mathbf{y}_j$. Then, we have

$$\begin{aligned}
\sigma_{i,j} &= \left\{\mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \frac{K_i}{C + \|\mathbf{x} - \mathbf{y}_i\|^2} > \frac{K_j}{C + \|\mathbf{x} - \mathbf{y}_j\|^2}\right\} \\
&\qquad\qquad [\text{from } (4.17), (4.16), (4.14), (4.23), (4.24)] \qquad (4.25) \\
&= \left\{\mathbf{x} \in \mathcal{S} : \|\mathbf{x} - (1 - \gamma_{i,j})\,\mathbf{y}_i - \gamma_{i,j}\mathbf{y}_j\|^2 \underset{K_i - K_j < 0}{\overset{K_i - K_j > 0}{\gtrless}} \gamma_{i,j}(\gamma_{i,j} - 1)\,\|\mathbf{y}_i - \mathbf{y}_j\|^2 - C\right\} \\
&\qquad\qquad [\text{from } (4.24)] \qquad (4.26)
\end{aligned}$$

where the notation $\underset{K_i - K_j < 0}{\overset{K_i - K_j > 0}{\gtrless}}$ indicates that "$>$" should be used if $K_i - K_j > 0$, and "$<$" if $K_i - K_j < 0$. By (4.26), the set $\sigma_{i,j}$ is either the interior or the exterior of a circle (depending on the sign of $K_i - K_j$). Applying (4.22) completes the proof. ■

Figure 4.1a shows the optimal sensor regions $\sigma_1, \sigma_2, \sigma_3$, and σ_4, for $N = 4$ randomly placed amplify-and-forward relays with AWGN channels and system parameter values $G = 65$ dB, $f_0 = 900$ MHz, and $C = 1$.

[8]Numerical results confirm that (4.23) is a close approximation of (4.2) for our parameters of interest.

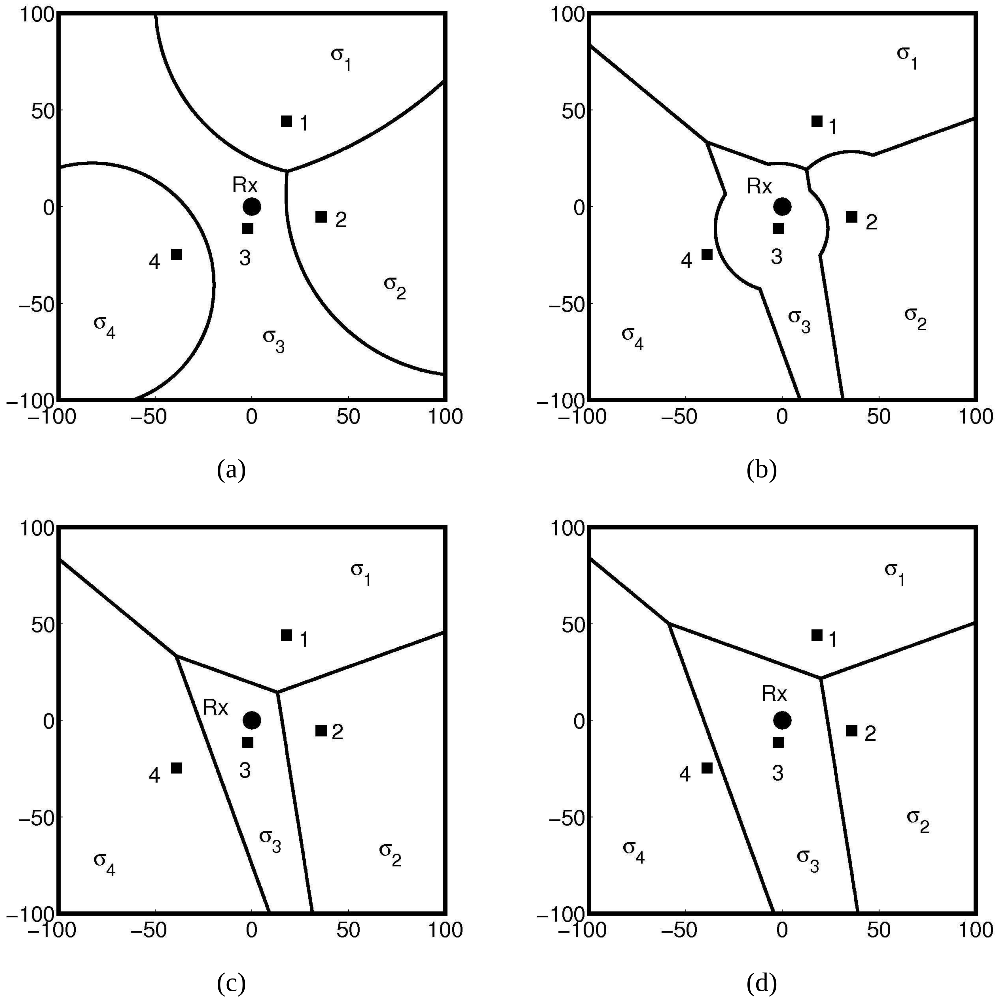

Figure 4.1: Sensor regions σ_1, σ_2, σ_3, and σ_4 for 4 randomly placed relays. Each relay $i \in \{1, 2, 3, 4\}$ is denoted by a filled square labeled i, while the receiver is denoted by a filled circle labeled Rx. Sensors are distributed as a square grid over ± 100 meters in each dimension. The sensor regions are either optimal or asymptotically-optimal as described in (a) Theorem 4.4.1 (amplify-and-forward relays and AWGN channels), (b) Theorem 4.4.4 (decode-and-forward relays and AWGN channels with high E_{Tx}/N_0 and E_i/N_0), (c) Theorem 4.4.6 (amplify-and-forward relays and Rayleigh fading channels with high E_{Tx}/N_0), and (d) Theorem 4.4.8 (decode-and-forward relays and Rayleigh fading channels) with high E_{Tx}/N_0 and E_i/N_0).

Decode-and-Forward with AWGN Channels

Lemma 4.4.2 (e.g., see [25, pp. 82 – 83], [24, pp. 37 – 39])**.** *For all $x > 0$,*

$$\left(1 - \frac{1}{x^2}\right)\left(\frac{e^{-x^2/2}}{\sqrt{2\pi}x}\right) \le Q(x) \le \frac{e^{-x^2/2}}{\sqrt{2\pi}x}.$$

Lemma 4.4.3. *Let* $\epsilon > 0$ *and*

$$L_{x,y} = \frac{Q(\sqrt{x}) + Q(\sqrt{y}) - 2Q(\sqrt{x})\,Q(\sqrt{y})}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}.$$

Then, $1 - \epsilon \leq L_{x,y} \leq 2$ *for* x *and* y *sufficiently large.*

Proof. For the lower bound, we have

$$\begin{aligned} L_{x,y} &\geq \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}} - \frac{\frac{e^{-x/2}}{x\sqrt{2\pi x}} + \frac{e^{-y/2}}{y\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right) \\ &\qquad\qquad \text{[from Lemma 4.4.2]} \\ &\geq 1 - \frac{1}{\min(x,y)} - \left(\frac{e^{-\max(x,y)/2}}{\max(x,y)\sqrt{\max(x,y)}}\right)\left(\frac{\sqrt{\min(x,y)}}{e^{-\min(x,y)/2}}\right) \\ &\qquad - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right) \qquad \text{[for } x, y > 1\text{]} \\ &\geq 1 - \epsilon. \qquad \text{[for } x, y \text{ sufficiently large]} \end{aligned}$$

For the upper bound, we have

$$\begin{aligned} L_{x,y} &\leq \frac{\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right) + \left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right) - 2\left(1 - \frac{1}{x}\right)\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right)\left(1 - \frac{1}{y}\right)\left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}{\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} \qquad \text{[from Lemma 4.4.2]} \\ &\leq \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} \qquad \text{[for } x, y > 1\text{]} \\ &\leq 2. \end{aligned}$$

■

Theorem 4.4.4. *Consider a sensor network with decode-and-forward relays and AWGN channels, and, for all relays* i, *let* $E_i/N_0 \to \infty$ *and* $E_{\mathrm{Tx}}/N_0 \to \infty$ *such that* $(E_i/N_0)/(E_{\mathrm{Tx}}/N_0)$ *has a limit. Then, the internal boundary of each optimal sensor region consists asymptotically of circular arcs and line segments.*

Proof. As an approximation to $P_e^{(\mathbf{x},i,\mathrm{Rx})}$ given in (4.19), define

$\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}$

$$= \frac{1}{\sqrt{2\pi}} \cdot \max\left(\frac{1}{\sqrt{\mathrm{SNR}^{(\mathbf{x},i)}}} \exp\left\{ -\frac{SNR^{(\mathbf{x},i)}}{2} \right\}, \frac{1}{\sqrt{\mathrm{SNR}^{(i,\mathrm{Rx})}}} \exp\left\{ -\frac{SNR^{(i,\mathrm{Rx})}}{2} \right\} \right). \tag{4.27}$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \dfrac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Let $\epsilon > 0$. Then, using Lemma 4.4.3, it can be shown that

$$1 - \epsilon \le \alpha_i \le 2. \tag{4.28}$$

We will now show that $\sigma_{i,j}$, given by (4.21), is a finite intersection of unions of certain sets $\rho_{i,j}^{(k)}$ for $k = 1, \ldots, 4$, where each such set has circular and/or linear boundaries.

For each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \ne j$, define

$$\begin{aligned} \rho_{i,j}^{(1)} &= \left\{ \mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} > \mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)} \right\} \\ &= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2F_2}{C + \|\mathbf{x} - \mathbf{y}_i\|^2} + \frac{N_0}{E_{\mathrm{Tx}}} \ln\left(\frac{\alpha_j}{\alpha_i}\right) + \frac{N_0}{E_{\mathrm{Tx}}} \ln\left(\frac{C + \|\mathbf{x} - \mathbf{y}_j\|^2}{C + \|\mathbf{x} - \mathbf{y}_i\|^2} \right) \right. \\ &\qquad \left. > \frac{2F_2}{C + \|\mathbf{x} - \mathbf{y}_j\|^2} \right\}. \qquad \text{[from (4.10), (4.23)]} \end{aligned}$$

The set $\mathcal{S}$ is bounded, so, using (4.28), as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, $\rho_{i,j}^{(1)} \to \left\{ \mathbf{x} \in \mathcal{S} : \|\mathbf{x} - \mathbf{y}_j\|^2 > \|\mathbf{x} - \mathbf{y}_i\|^2 \right\}$ which has a linear internal boundary.

Also, for each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \ne j$, define

$$\begin{aligned} \rho_{i,j}^{(2)} &= \left\{ \mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} > \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})} \right\} \\ &= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2F_2}{C + \|\mathbf{x} - \mathbf{y}_i\|^2} \right. \\ &\qquad > \frac{2F_2}{C + \|\mathbf{y}_j\|^2} \cdot \frac{E_j/N_0}{E_{\mathrm{Tx}}/N_0} + \frac{N_0}{E_{\mathrm{Tx}}} \ln\left(\frac{C + \|\mathbf{x} - \mathbf{y}_i\|^2}{C + \|\mathbf{y}_j\|^2} \cdot \frac{E_j/N_0}{E_{\mathrm{Tx}}/N_0} \right) \\ &\qquad \left. + \frac{N_0}{E_{\mathrm{Tx}}} \ln\left(\frac{\alpha_i}{\alpha_j}\right) \right\}. \qquad \text{[from (4.10), (4.23)]} \end{aligned} \tag{4.29}$$

In the cases that follow, we will show that, asymptotically, $\rho_{i,j}^{(2)}$ either contains all of the sensors, none of the sensors, or the subset of sensors in the interior of a circle.

Case 1: $(E_j/N_0)/(E_{\mathrm{Tx}}/N_0) \to \infty$.

The set $\mathcal{S}$ is bounded and, by (4.28), $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded. Therefore, the limit of the right-hand side of the inequality in (4.29) is infinity. Thus, $\rho_{i,j}^{(2)} \to \emptyset$.

Case 2: $(E_j/N_0)/(E_{\text{Tx}}/N_0) \to G_j$ for some $G_j \in (0, \infty)$.

Since $\mathcal{S}$ is bounded and $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded, we have $\rho_{i,j}^{(2)} \to \left\{\mathbf{x} \in \mathcal{S} : \|\mathbf{x} - \mathbf{y}_i\|^2 < \frac{C+\|\mathbf{y}_j\|^2}{G_j} - C\right\}$ which has a circular internal boundary.

Case 3: $(E_j/N_0)/(E_{\text{Tx}}/N_0) \to 0$.

Since $\mathcal{S}$ is bounded and $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded, the limit of the right-hand side of the inequality in (4.29) is 0. Thus, since $F_2 > 0$, we have $\rho_{i,j}^{(2)} \to \mathcal{S}$.

Also, for each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \neq j$, define

$$\rho_{i,j}^{(3)} = \left\{\mathbf{x} \in \mathcal{S} : \text{SNR}^{(i,\text{Rx})} - 2\ln\alpha_i + \ln \text{SNR}^{(i,\text{Rx})} > \text{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \text{SNR}^{(\mathbf{x},j)}\right\}.$$

Observing the symmetry between $\rho_{i,j}^{(3)}$ and $\rho_{i,j}^{(2)}$, we have that as $E_{\text{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, $\rho_{i,j}^{(3)}$ becomes either empty, all of $\mathcal{S}$, or the exterior of a circle.

Also, for each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \neq j$, define

$$\begin{aligned}
\rho_{i,j}^{(4)} &= \Big\{\mathbf{x} \in \mathcal{S} : \text{SNR}^{(i,\text{Rx})} - 2\ln\alpha_i + \ln \text{SNR}^{(i,\text{Rx})} \\
&\qquad > \text{SNR}^{(j,\text{Rx})} - 2\ln\alpha_j + \ln \text{SNR}^{(j,\text{Rx})}\Big\} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \frac{2E_iF_2}{N_0\left(C + \|\mathbf{y}_i\|^2\right)} - \ln\alpha_i + \ln\left(\frac{2E_iF_2}{N_0\left(C + \|\mathbf{y}_i\|^2\right)}\right)\right. \\
&\qquad \left. > \frac{2E_jF_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)} - \ln\alpha_j + \ln\left(\frac{2E_jF_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)}\right)\right\}. \\
&\qquad\qquad \text{[from (4.10), (4.23)]}
\end{aligned}$$

Using (4.28), as $E_{\text{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, we have $\rho_{i,j}^{(4)} \to \mathcal{S}$ or $\emptyset$.

Then, we have

$$\begin{aligned}
\sigma_{i,j} &= \left\{\mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\text{Rx})} < P_e^{(\mathbf{x},j,\text{Rx})}\right\} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \alpha_i \hat{P}_e^{(\mathbf{x},i,\text{Rx})} < \alpha_j \hat{P}_e^{(\mathbf{x},j,\text{Rx})}\right\} \\
&= \Big\{\mathbf{x} \in \mathcal{S} : \min\Big(\text{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \text{SNR}^{(\mathbf{x},i)}, \\
&\qquad\qquad \text{SNR}^{(i,\text{Rx})} - 2\ln\alpha_i + \ln \text{SNR}^{(i,\text{Rx})}\Big) \\
&\qquad > \min\Big(\text{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \text{SNR}^{(\mathbf{x},j)}, \\
&\qquad\qquad \text{SNR}^{(j,\text{Rx})} - 2\ln\alpha_j + \ln \text{SNR}^{(j,\text{Rx})}\Big)\Big\} \\
&\qquad \text{[for } E_{\text{Tx}}/N_0, E_i/N_0, E_j/N_0 \text{ sufficiently large]} \qquad \text{[from (4.27)]}
\end{aligned}$$

$$= \left(\rho_{i,j}^{(1)} \cup \rho_{i,j}^{(2)}\right) \cap \left(\rho_{i,j}^{(3)} \cup \rho_{i,j}^{(4)}\right). \tag{4.30}$$

Thus, combining the asymptotic results for $\rho_{i,j}^{(1)}, \rho_{i,j}^{(2)}, \rho_{i,j}^{(3)}$, and $\rho_{i,j}^{(4)}$, as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, the internal boundary of $\sigma_{i,j}$ consists of circular arcs and line segments. Applying (4.22) completes the proof. ■

Figure 4.1b shows the asymptotically-optimal sensor regions $\sigma_1, \sigma_2, \sigma_3$, and σ_4, for $N = 4$ randomly placed decode-and-forward relays with AWGN channels and system parameter values $C = 1$, $E_{\mathrm{Rx}}/N_0|_{d=50 \text{ m}} = 5$ dB, and $E_i/N_0 = 2E_{\mathrm{Tx}}/N_0$ for all relays $\mathbf{y}_i$.

Amplify-and-Forward with Rayleigh Fading Channels

Lemma 4.4.5. *For* $0 < z < 1$,

$$\left(\frac{1}{z\Gamma\left(\frac{3}{2}\right)}\right)\left(1-\sqrt{z}\right)\exp\left\{-\frac{\sqrt{z}\,(1-\sqrt{z})^2}{2-\sqrt{z}}\right\} \le U\left(\frac{3}{2}, 2, z\right) \le \frac{1}{z\Gamma\left(\frac{3}{2}\right)}.$$

Proof. For the upper bound, we have

$$U\left(\frac{3}{2}, 2, z\right) = \frac{1}{\Gamma\left(\frac{3}{2}\right)}\int_0^\infty \sqrt{\frac{t}{1+t}} \cdot e^{-zt}dt \le \frac{1}{\Gamma\left(\frac{3}{2}\right)}\int_0^\infty e^{-zt}dt = \frac{1}{z\Gamma\left(\frac{3}{2}\right)}.$$

For the lower bound, we have

$$\begin{aligned}
U\left(\frac{3}{2}, 2, z\right) &\ge \frac{1}{\Gamma\left(\frac{3}{2}\right)}\int_{\frac{(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}}^\infty \sqrt{\frac{t}{1+t}} \cdot e^{-zt}dt && [\text{since } 0 < z < 1] \\
&\ge \frac{1}{\Gamma\left(\frac{3}{2}\right)}\int_{\frac{(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}}^\infty (1-\sqrt{z})e^{-zt}dt && [\text{since } 0 < z < 1] \\
&= \frac{1}{z\Gamma\left(\frac{3}{2}\right)}\left(1-\sqrt{z}\right)\exp\left\{-\frac{\sqrt{z}(1-\sqrt{z})^2}{2-\sqrt{z}}\right\}.
\end{aligned}$$

■

We define the *nearest-neighbor region* of a relay $\mathbf{y}_i$ to be

$$\{\mathbf{x} \in \mathcal{S} : \forall j,\ \|\mathbf{x} - \mathbf{y}_i\| < \|\mathbf{x} - \mathbf{y}_j\|\}$$

where ties (i.e., $\|\mathbf{x} - \mathbf{y}_i\| = \|\mathbf{x} - \mathbf{y}_j\|$) are broken arbitrarily. The interiors of these regions are convex polygons intersected with $\mathcal{S}$.

Theorem 4.4.6. *Consider a sensor network with amplify-and-forward relays and Rayleigh fading channels, and let $E_{\mathrm{Tx}}/N_0 \to \infty$. Then, each optimal sensor region is asymptotically equal to the corresponding relay's nearest-neighbor region.*

Proof. As an approximation to $P_e^{(\mathbf{x},i,\mathrm{Rx})}$ given in (4.15), define

$$\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} = \frac{1}{2} - \left(\frac{D_i \sqrt{\pi} N_0/E_{\mathrm{Tx}}}{8\sigma \left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)^{3/2}} \right) \left(\frac{2\sigma^2 \left(\sigma^2 + B_i N_0/E_{\mathrm{Tx}}\right)}{\Gamma(3/2) \cdot D_i N_0/E_{\mathrm{Tx}}} \right) \tag{4.31}$$

$$= \frac{1}{2} - \frac{1}{2} \left(1 + \frac{1}{2\sigma^2 L_{\mathbf{x},i} E_{\mathrm{Tx}}/N_0} \right)^{-1/2}. \qquad \text{[from (4.14)]} \tag{4.32}$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \dfrac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Using Lemma 4.4.5, it can be shown that

$$\lim_{E_{\mathrm{Tx}}/N_0 \to \infty} \alpha_i = 1. \tag{4.33}$$

Let

$$Z_k = \frac{1}{2\sigma^2 L_{\mathbf{x},k}}; \qquad g_k\left(\frac{N_0}{E_{\mathrm{Tx}}}\right) = \sqrt{1 + \frac{Z_k N_0}{E_{\mathrm{Tx}}}} - 1 = \left(\frac{Z_k}{2}\right) \frac{N_0}{E_{\mathrm{Tx}}} + \mathcal{O}\left(\left(\frac{N_0}{E_{\mathrm{Tx}}}\right)^2 \right) \tag{4.34}$$

where the second equality in the expression for g_k is obtained using a Taylor series. Then,

$$\begin{aligned}
\sigma_{i,j} &= \left\{ \mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})} \right\} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \alpha_i \hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} < \alpha_j \hat{P}_e^{(\mathbf{x},j,\mathrm{Rx})} \right\} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i \left(\sqrt{1 + \frac{Z_i N_0}{E_{\mathrm{Tx}}}} - 1 \right) \sqrt{1 + \frac{Z_j N_0}{E_{\mathrm{Tx}}}}}{\alpha_j \left(\sqrt{1 + \frac{Z_j N_0}{E_{\mathrm{Tx}}}} - 1 \right) \sqrt{1 + \frac{Z_i N_0}{E_{\mathrm{Tx}}}}} < 1 \right\} \qquad \text{[from (4.32), (4.34)]} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i}{\alpha_j} \cdot \frac{\frac{1}{4\sigma^2 L_{\mathbf{x},i}} + \mathcal{O}\left(\frac{N_0}{E_{\mathrm{Tx}}}\right)}{\frac{1}{4\sigma^2 L_{\mathbf{x},j}} + \mathcal{O}\left(\frac{N_0}{E_{\mathrm{Tx}}}\right)} \cdot \sqrt{\frac{1 + \frac{N_0/E_{\mathrm{Tx}}}{2\sigma^2 L_{\mathbf{x},j}}}{1 + \frac{N_0/E_{\mathrm{Tx}}}{2\sigma^2 L_{\mathbf{x},i}}}} < 1 \right\}. \qquad \text{[from (4.34)]}
\end{aligned} \tag{4.35}$$

Since $\mathcal{S}$ is bounded, we have, for $E_{\mathrm{Tx}}/N_0 \to \infty$, that

$$\sigma_{i,j} \to \left\{ \mathbf{x} \in \mathcal{S} : \|\mathbf{x} - \mathbf{y}_j\| > \|\mathbf{x} - \mathbf{y}_i\| \right\}. \qquad \text{[from (4.35), (4.33), (4.23)]} \tag{4.36}$$

Thus, for $E_{\text{Tx}}/N_0 \to \infty$, the internal boundary of $\sigma_{i,j}$ becomes the line equidistant from $\mathbf{y}_i$ and $\mathbf{y}_j$. Applying (4.22) completes the proof. ■

Figure 4.1c shows the asymptotically-optimal sensor regions $\sigma_1, \sigma_2, \sigma_3$, and σ_4, for $N = 4$ randomly placed amplify-and-forward relays with Rayleigh fading channels.

Decode-and-Forward with Rayleigh Fading Channels

Lemma 4.4.7. *Let*

$$L_{x,y} = \frac{1 - \left(1 + \frac{2}{x}\right)^{-1/2} \left(1 + \frac{2}{y}\right)^{-1/2}}{\frac{1}{x} + \frac{1}{y}}.$$

Then, $\lim_{x,y\to\infty} L_{x,y} = 1$.

Proof. We have

$$\begin{aligned} & 1 + \frac{1}{2}\epsilon - \frac{1}{8}\epsilon^2 \le (1+\epsilon)^{1/2} \le 1 + \frac{1}{2}\epsilon && \text{[from a Taylor series]} \\ \therefore\ & \left(\frac{xy}{x+y}\right) \frac{\left(x - \frac{1}{2}\right)\left(y^2 + y - \frac{1}{2}\right) + x^2\left(y - \frac{1}{2}\right)}{\left(x^2 + x - \frac{1}{2}\right)\left(y^2 + y - \frac{1}{2}\right)} \le L_{x,y} \\ & \quad \le \left(\frac{x+y+1}{x+y}\right)\left(\frac{x}{x+1}\right)\left(\frac{y}{y+1}\right) \\ \therefore\ & \left(\frac{x-1}{x+1}\right)\left(\frac{y-1}{y+1}\right)\left(\frac{x+y+3}{x+y}\right) \le L_{x,y} \le \left(\frac{x+y+1}{x+y}\right)\left(\frac{x}{x+1}\right)\left(\frac{y}{y+1}\right). \\ & && \text{[for } x, y \text{ sufficiently large]} \end{aligned}$$

Now taking the limit as $x \to \infty$ and $y \to \infty$ (in any manner) gives $L_{x,y} \to 1$. ■

Theorem 4.4.8. *Consider a sensor network with decode-and-forward relays and Rayleigh fading channels, and, for all relays i, let $E_i/N_0 \to \infty$ and $E_{\text{Tx}}/N_0 \to \infty$ such that $(E_i/N_0)/(E_{\text{Tx}}/N_0)$ has a limit. Then, the internal boundary of each optimal sensor region is asymptotically piecewise linear.*

Proof. As an approximation to $P_e^{(\mathbf{x},i,\text{Rx})}$ given in (4.18), define

$$\hat{P}_e^{(\mathbf{x},i,\text{Rx})} = \frac{1/2}{\text{SNR}^{(\mathbf{x},i)}} + \frac{1/2}{\text{SNR}^{(i,\text{Rx})}}. \tag{4.37}$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \dfrac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Using Lemma 4.4.7, it can be shown that

$$\lim_{\substack{E_{\mathrm{Tx}}/N_0 \to \infty, \\ E_i/N_0 \to \infty}} \alpha_i = 1. \tag{4.38}$$

Then, we have

$$\begin{aligned}
\sigma_{i,j} &= \left\{ \mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})} \right\} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \alpha_i \hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} < \alpha_j \hat{P}_e^{(\mathbf{x},j,\mathrm{Rx})} \right\} \\
&= \Bigg\{ \mathbf{x} \in \mathcal{S} : 2 \left\langle \mathbf{x}, \alpha_j \mathbf{y}_j - \alpha_i \mathbf{y}_i \right\rangle \\
&\qquad < \alpha_j \left(C + \|\mathbf{y}_j\|^2 \right) \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} - \alpha_i \left(C + \|\mathbf{y}_i\|^2 \right) \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0} \\
&\qquad + (\alpha_j - \alpha_i) \|\mathbf{x}\|^2 + \alpha_j \|\mathbf{y}_j\|^2 - \alpha_i \|\mathbf{y}_i\|^2 \Bigg\} . \\
&\qquad\qquad [\text{from } (4.37), (4.8), (4.23)]
\end{aligned} \tag{4.39}$$

Now, for any relay $\mathbf{y}_k$, let $G_k = \lim_{\substack{E_{\mathrm{Tx}}/N_0 \to \infty, \\ E_k/N_0 \to \infty}} \dfrac{E_k/N_0}{E_{\mathrm{Tx}}/N_0}$. Using (4.38), Table 4.1 considers the cases of G_i and G_j being zero, infinite, or finite non-zero; for all such possibilities, the internal boundary of $\sigma_{i,j}$ is linear. Applying (4.22) completes the proof. ■

Note that if, for all relays $\mathbf{y}_i$, E_i is a constant and $G_i = \infty$, then each optimal sensor region is asymptotically equal to the corresponding relay's nearest-neighbor regions, as was the case for amplify-and-forward relays and Rayleigh fading channels. In addition, we note that, while Theorem 4.4.8 considers the asymptotic case, we have empirically observed that the internal boundary of each optimal sensor region consists of line segments for a wide range of moderate parameter values.

Table 4.1: Asymptotic properties of $\sigma_{i,j}$ for decode-and-forward relays and Rayleigh fading channels.

G_j	G_i	$\sigma_{i,j}$
non-zero	non-zero	linear internal boundary
non-zero	0	$\emptyset$
0	non-zero	$\mathcal{S}$
0	0	linear internal boundary or $\emptyset$ or $\mathcal{S}$

Figure 4.1d shows the asymptotically-optimal sensor regions $\sigma_1, \sigma_2, \sigma_3$, and σ_4, for $N = 4$ randomly placed decode-and-forward relays with Rayleigh fading channels and system parameter values $C = 1$, $E_{\mathrm{Rx}}/N_0|_{d=50 \text{ m}} = 5$ dB, and $E_i/N_0 = 2E_{\mathrm{Tx}}/N_0$ for all relays $\mathbf{y}_i$.

4.5 Numerical Results for the Relay Placement Algorithm

The relay placement algorithm was implemented for both amplify-and-forward and decode-and-forward relays. The sensors were placed uniformly in a square of sidelength 100 m. For decode-and-forward and all relays $\mathbf{y}_i$, the energy E_i was set to a constant which equalized the total output power of all relays for both amplify-and-forward and decode-and-forward. Specific numerical values for system variables were $f_0 = 900$ MHz, $\sigma = \sqrt{2}/2$, $M = 10000$, and $C = 1$.

In order to use the relay placement algorithm to produce good relay locations and sensor-relay assignments, we ran the algorithm 10 times. Each such run was initiated with a different random set of relay locations (uniformly distributed on the square $\mathcal{S}$) and used the sensor-averaged probability of error given in (4.20). For each of the 10 runs completed, 1000 simulations were performed with Rayleigh fading and diversity (selection combining) at the receiver. Different realizations of the fade values for the sensor network channels were chosen for each of the 1000 simulations. Of the 10 runs, the relay locations and sensor-relay assignments of the run with the lowest average probability of error over the 1000 simulations was chosen.

Figure 4.2 gives the algorithm output for 2, 3, 4, and 12 decode-and-forward relays with $E_{\mathrm{Rx}}/N_0|_{d=50 \text{ m}} = 10$ dB, $E_i = 100E_{\mathrm{Tx}}$, and using the exact error probability expressions. Relays are denoted by squares and the receiver is denoted by a circle at the origin. Boundaries between the optimal sensor regions are shown. For 2, 3, and 4 relays a symmetry is present, with each relay being responsible for approximately the same number of sensors. A symmetry is also present for 12 relays; here, however, eight relays are responsible for approximately the same number of sensors, and the remaining four relays are located near the corners of $\mathcal{S}$ to assist in transmissions experiencing the largest path loss due to distance. Since the relays transmit at higher energies than the sensors, the

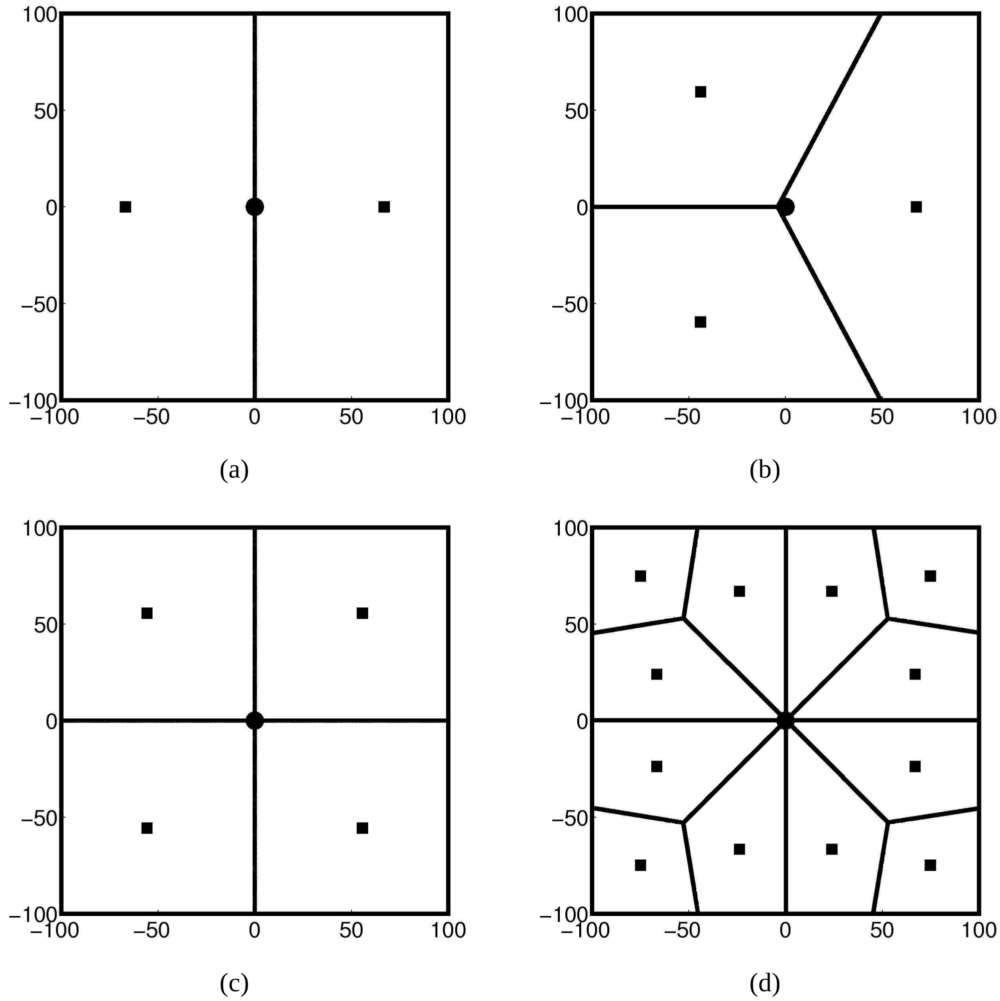

Figure 4.2: Optimal sensor regions output by the algorithm for decode-and-forward relays and fading channels with $E_i = 100E_{\mathrm{Tx}}$, and $E_{\mathrm{Rx}}/N_0|_{d=50\ \mathrm{m}} = 10$ dB. Relays are denoted by squares and the receiver is located at $(0,0)$. Sensors are distributed as a square grid over ± 100 meters in each dimension. The number of relays is (a) $N = 2$, (b) $N = 3$, (c) $N = 4$, and (d) $N = 12$.

probability of detection error is reduced by reducing path loss before a relay rebroadcasts a sensor's signal, rather than after the relay rebroadcasts the signal (even at the expense of possibly greater path loss from the relay to the receiver). Thus, some sensors actually transmit "away" from the receiver to their associated relay. The asymptotically-optimal sensor regions closely matched those for the exact error probability expressions, which is expected due to the large value selected for E_i. In addition, the results for amplify-and-for-

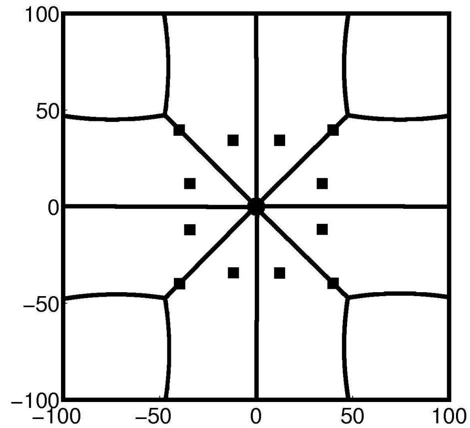

Figure 4.3: Optimal sensor regions $\sigma_1, \ldots, \sigma_{12}$ output by the algorithm for decode-and-forward relays and fading channels with $N = 12$, $E_i = 1.26E_{\mathrm{Tx}}$, and $E_{\mathrm{Rx}}/N_0|_{d=50\text{ m}} = 5$ dB.

ward relays were quite similar, with the relays lying closer to the corners of $\mathcal{S}$ for the 2 and 3 relay cases, and the corner regions displaying slightly curved boundaries for 12 relays. With the exception of this curvature, the asymptotic regions closely matched those from the exact error probability expressions. This similarity between decode-and-forward and amplify-and-forward relays is expected due to the large value selected for E_i.

Figures 4.3 and 4.4 give the algorithm output for 12 decode-and-forward and amplify-and-forward relays, respectively, with $E_{\mathrm{Rx}}/N_0|_{d=50\text{ m}} = 5$ dB, $E_i = 1.26E_{\mathrm{Tx}}$, and using the exact error probability expressions. For decode-and-forward relays, the results are similar to those in Figure 4.3; however the relays are located much closer to the receiver due to their decreased transmission energy, and the corner regions of $\mathcal{S}$ exhibit slightly curved boundaries. For amplify-and-forward relays, the relays are located much closer to the corners since, with lower gain, the relays are less effective and thus primarily assist those sensors with the largest path loss.

The maximum, average, and median of the sensor probabilities of error for all of the above figures are given in Table 4.2. The sensor error probability is lowest for sensors that are closest to the relays, and increases with distance.

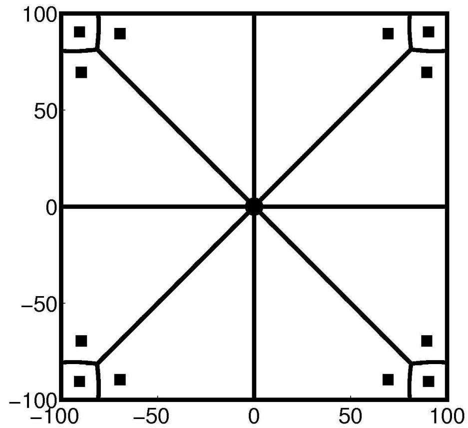

Figure 4.4: Optimal sensor regions $\sigma_1, \ldots, \sigma_{12}$ output by the algorithm for amplify-and-forward relays and fading channels with $N = 12$, $G = 56$ dB, and $E_{\mathrm{Rx}}/N_0|_{d=50\ \mathrm{m}} = 5$ dB.

Table 4.2: Sensor probability of error values.

Figure	Max. P_e	Avg. P_e	Median P_e
4.2a	$7.3 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$
4.2b	$6.9 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$7.2 \cdot 10^{-3}$
4.2c	$3.3 \cdot 10^{-2}$	$7.0 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$
4.2d	$1.4 \cdot 10^{-2}$	$2.8 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$
4.3	$2.0 \cdot 10^{-1}$	$6.2 \cdot 10^{-2}$	$5.6 \cdot 10^{-2}$
4.4	$1.7 \cdot 10^{-1}$	$9.9 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$

4.6 Conclusions

This paper presented an algorithm for amplify-and-forward and decode-and-forward relay placement and sensor assignment in wireless sensor networks that attempts to minimize the average probability of error. Communications were modeled using path loss, fading, AWGN, and diversity combining. We determined the geometric shapes of regions for which sensors would be optimally assigned to the same relay (for a given set of relay locations), in some instances for the asymptotic case of the ratios of the transmission energies to the noise power spectral density growing without bound. Numerical results showing the algorithm output were presented. The asymptotic regions were seen to closely match the regions obtained using exact expressions.

A number of extensions to the relay placement algorithm could be incorporated to enhance the system model. Some such enhancements are multi-hop relay paths, more sophisticated diversity combining, power constraints, sensor priorities, and sensor information correlation.

4.7 Acknowledgment

With the exception of the appendix, the text of this chapter, in full, has been submitted for publication as Jillian Cannons, Laurence B. Milstein, and Kenneth Zeger, “An Algorithm for Wireless Relay Placement,” *IEEE Transactions on Wireless Communications*, submitted August 4, 2008.

Appendix

This appendix contains expanded versions of proofs in this chapter.

Expanded Proof of Theorem 4.4.1. For any distinct relays $\mathbf{y}_i$ and $\mathbf{y}_j$, let

$$K_i = \frac{1}{G^2 F_2 + C + \|\mathbf{y}_i\|^2} \qquad \gamma_{i,j} = \frac{K_i}{K_i - K_j}. \tag{4.40}$$

Note that for fixed gain G, $K_i \neq K_j$ since we assume $\mathbf{y}_i \neq \mathbf{y}_j$. Then, we have

$$\begin{aligned}
&\sigma_{i,j} \\
&= \left\{\mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i,\mathrm{Rx})} > \mathrm{SNR}^{(\mathbf{x},j,\mathrm{Rx})}\right\} && \text{[from (4.17)]} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \frac{1}{B_i + D_i} > \frac{1}{B_j + D_j}\right\} && \text{[from (4.16)]} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \frac{1}{1/(2L_{\mathbf{x},i}) + 1/(2G^2 L_{\mathbf{x},i} L_{i,\mathrm{Rx}})} \right. \\
&\qquad\qquad \left. > \frac{1}{1/(2L_{\mathbf{x},j}) + 1/(2G^2 L_{\mathbf{x},j} L_{j,\mathrm{Rx}})}\right\} && \text{[from (4.14)]} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \frac{L_{\mathbf{x},i} L_{i,\mathrm{Rx}}}{G^2 L_{i,\mathrm{Rx}} + 1} > \frac{L_{\mathbf{x},j} L_{j,\mathrm{Rx}}}{G^2 L_{j,\mathrm{Rx}} + 1}\right\}
\end{aligned}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{1}{G^2 F_2 (C + d_{\mathbf{x},i}^2) + (C + d_{\mathbf{x},i}^2)(C + d_{i,\mathrm{Rx}}^2)} \right.$$
$$\left. > \frac{1}{G^2 F_2 (C + d_{\mathbf{x},j}^2) + (C + d_{\mathbf{x},j}^2)(C + d_{j,\mathrm{Rx}}^2)} \right\} \qquad \text{[from (4.23)]}$$
$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{K_i}{C + \|\mathbf{x} - \mathbf{y}_i\|^2} > \frac{K_j}{C + \|\mathbf{x} - \mathbf{y}_j\|^2} \right\} \qquad \text{[from (4.40)]}$$
$$= \left\{ \mathbf{x} \in \mathcal{S} : \|\mathbf{x} - (1 - \gamma_{i,j})\,\mathbf{y}_i - \gamma_{i,j}\mathbf{y}_j\|^2 \underset{K_i - K_j < 0}{\overset{K_i - K_j > 0}{\gtrless}} \gamma_{i,j}\,(\gamma_{i,j} - 1)\,\|\mathbf{y}_i - \mathbf{y}_j\|^2 - C \right\}$$
$$\text{[from (4.40)]}$$

where the notation $\underset{K_i - K_j < 0}{\overset{K_i - K_j > 0}{\gtrless}}$ indicates that "$>$" should be used if $K_i - K_j > 0$, and "$<$" if $K_i - K_j < 0$. Note that the description of $\sigma_{i,j}$ given in (4.26) is either the interior or the exterior of a circle (depending on the sign of $K_i - K_j$). Applying (4.22) completes the proof. ∎

Extended Proof of Lemma 4.4.3. For the lower bound we have

$$L_{x,y} \geq \frac{\left(1 - \frac{1}{x}\right)\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right) + \left(1 - \frac{1}{y}\right)\left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right) - 2\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right)\left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} \qquad \text{[from Lemma 4.4.2]}$$
$$= \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - \frac{\frac{e^{-x/2}}{x\sqrt{2\pi x}} + \frac{e^{-y/2}}{y\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - \frac{2\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right)\left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}$$
$$= \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - \frac{\frac{e^{-x/2}}{x\sqrt{2\pi x}} + \frac{e^{-y/2}}{y\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)$$
$$\geq \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}} - \frac{\frac{e^{-x/2}}{x\sqrt{2\pi x}} + \frac{e^{-y/2}}{y\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)$$
$$= 1 - \frac{\frac{e^{-x/2}}{x\sqrt{2\pi x}} + \frac{e^{-y/2}}{y\sqrt{2\pi y}}}{\max\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)$$
$$\geq 1 - \frac{1}{\min(x,y)} - \left(\frac{e^{-\max(x,y)/2}}{\max(x,y)\sqrt{\max(x,y)}}\right)\left(\frac{\sqrt{\min(x,y)}}{e^{-\min(x,y)/2}}\right)$$
$$- 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right) \qquad \text{[for } x, y > 1\text{]}$$

$$
\begin{aligned}
&= 1 - \frac{1}{\min(x,y)} - \frac{e^{-(\max(x,y)-\min(x,y))/2}}{\max(x,y)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)\\
&\geq 1 - \frac{1}{\min(x,y)} - \frac{e^{0}}{\max(x,y)} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)\\
&= 1 - \frac{1}{x} - \frac{1}{y} - 2\min\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)\\
&\geq 1 - \epsilon. \qquad \text{[for } x, y \text{ sufficiently large]}
\end{aligned}
$$

For the upper bound we have

$$
\begin{aligned}
L_{x,y} &\leq \frac{\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right) + \left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right) - 2\left(1-\frac{1}{x}\right)\left(\frac{e^{-x/2}}{\sqrt{2\pi x}}\right)\left(1-\frac{1}{y}\right)\left(\frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}{\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} \qquad \text{[from Lemma 4.4.2]}\\
&\leq \frac{\frac{e^{-x/2}}{\sqrt{2\pi x}} + \frac{e^{-y/2}}{\sqrt{2\pi y}}}{\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)} \qquad \text{[for } x, y > 1\text{]}\\
&\leq \frac{2\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}{\max\left(\frac{e^{-2/x}}{\sqrt{2\pi x}}, \frac{e^{-y/2}}{\sqrt{2\pi y}}\right)}\\
&= 2.
\end{aligned}
$$

■

Extended Proof of Theorem 4.4.4. As an approximation to $P_e^{(\mathbf{x},i,\mathrm{Rx})}$ given in (4.19), define

$$
\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} = \frac{1}{\sqrt{2\pi}} \cdot \max\left(\frac{1}{\sqrt{\mathrm{SNR}^{(\mathbf{x},i)}}} \exp\left\{-\frac{SNR^{(\mathbf{x},i)}}{2}\right\}, \frac{1}{\sqrt{\mathrm{SNR}^{(i,\mathrm{Rx})}}} \exp\left\{-\frac{SNR^{(i,\mathrm{Rx})}}{2}\right\}\right). \tag{4.41}
$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \frac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Let $\epsilon > 0$. Then, using Lemma 4.4.3, it can be shown that

$$
1 - \epsilon \leq \alpha_i \leq 2. \tag{4.42}
$$

We will now show that $\sigma_{i,j}$, given by (4.21), is a finite intersection of unions of certain sets $\rho_{i,j}^{(k)}$ for $k = 1, \dots, 4$, where each such set has circular and/or linear boundaries.

For each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \neq j$, define

$$\begin{aligned}
\rho_{i,j}^{(1)} &= \left\{ \mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} > \mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)} \right\} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_i\|^2\right)} - \ln\alpha_i + \ln\left(\frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_i\|^2\right)}\right) \right. \\
&\qquad \left. > \frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_j\|^2\right)} - \ln\alpha_j + \ln\left(\frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_j\|^2\right)}\right) \right\} \\
&\qquad\qquad \text{[from (4.10), (4.23)]} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2F_2}{C + \|\mathbf{x}-\mathbf{y}_i\|^2} + \frac{N_0}{E_{\mathrm{Tx}}}\ln\left(\frac{\alpha_j}{\alpha_i}\right) + \frac{N_0}{E_{\mathrm{Tx}}}\ln\left(\frac{C + \|\mathbf{x}-\mathbf{y}_j\|^2}{C + \|\mathbf{x}-\mathbf{y}_i\|^2}\right) \right. \\
&\qquad \left. > \frac{2F_2}{C + \|\mathbf{x}-\mathbf{y}_j\|^2} \right\}. \qquad \text{[from (4.10), (4.23)]}
\end{aligned}$$

The set $\mathcal{S}$ is bounded, so, using (4.42), as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, $\rho_{i,j}^{(1)} \to \left\{ \mathbf{x} \in \mathcal{S} : \frac{2F_2}{C+\|\mathbf{x}-\mathbf{y}_i\|^2} > \frac{2F_2}{C+\|\mathbf{x}-\mathbf{y}_j\|^2} \right\} = \left\{ \mathbf{x} \in \mathcal{S} : \|\mathbf{x}-\mathbf{y}_j\|^2 > \|\mathbf{x}-\mathbf{y}_i\|^2 \right\}$ which has a linear internal boundary.

Also, for each pair of relays $(\mathbf{y}_i, \mathbf{y}_j)$ with $i \neq j$, define

$$\begin{aligned}
\rho_{i,j}^{(2)} &= \left\{ \mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} > \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})} \right\} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_i\|^2\right)} - \ln\alpha_i + \ln\left(\frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C + \|\mathbf{x}-\mathbf{y}_i\|^2\right)}\right) \right. \\
&\qquad \left. > \frac{2E_jF_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)} - \ln\alpha_j + \ln\left(\frac{2E_jF_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)}\right) \right\} \\
&\qquad\qquad \text{[from (4.10), (4.23)]} \\
&= \left\{ \mathbf{x} \in \mathcal{S} : \frac{2F_2}{C + \|\mathbf{x}-\mathbf{y}_i\|^2} \right. \\
&\qquad > \frac{2F_2}{C + \|\mathbf{y}_j\|^2} \cdot \frac{E_j/N_0}{E_{\mathrm{Tx}}/N_0} + \frac{N_0}{E_{\mathrm{Tx}}}\ln\left(\frac{C + \|\mathbf{x}-\mathbf{y}_i\|^2}{C + \|\mathbf{y}_j\|^2} \cdot \frac{E_j/N_0}{E_{\mathrm{Tx}}/N_0}\right) \\
&\qquad\quad \left. + \frac{N_0}{E_{\mathrm{Tx}}}\ln\left(\frac{\alpha_i}{\alpha_j}\right) \right\}. \qquad \text{[from (4.10), (4.23)]} \qquad (4.43)
\end{aligned}$$

In the cases that follow, we will show that, asymptotically, $\rho_{i,j}^{(2)}$ either contains all of the sensors, none of the sensors, or the subset of sensors in the interior of a circle.

Case 1: $(E_j/N_0)/(E_{\mathrm{Tx}}/N_0) \to \infty$.

The set $\mathcal{S}$ is bounded and, by (4.42), $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded. Therefore, the limit of the right-hand side of the inequality in (4.43) is infinity. Thus, $\rho_{i,j}^{(2)} \to \emptyset$.

Case 2: $(E_j/N_0)/(E_{\mathrm{Tx}}/N_0) \to G_j$ for some $G_j \in (0,\infty)$.

Since $\mathcal{S}$ is bounded and $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded, we have

$$\begin{aligned}\rho_{i,j}^{(2)} &\to \left\{\mathbf{x}\in\mathcal{S} : \frac{2F_2}{C+\|\mathbf{x}-\mathbf{y}_i\|^2} > \frac{2F_2}{C+\|\mathbf{y}_j\|^2}\cdot G_j\right\}\\ &= \left\{\mathbf{x}\in\mathcal{S} : \frac{C+\|\mathbf{y}_j\|^2}{G_j} > C+\|\mathbf{x}-\mathbf{y}_i\|^2\right\}\\ &= \left\{\mathbf{x}\in\mathcal{S} : \|\mathbf{x}-\mathbf{y}_i\|^2 < \frac{C+\|\mathbf{y}_j\|^2}{G_j} - C\right\}\end{aligned}$$

which has a circular internal boundary.

Case 3: $(E_j/N_0)/(E_{\mathrm{Tx}}/N_0) \to 0$.

Since $\mathcal{S}$ is bounded and $\ln(\alpha_i/\alpha_j)$ is asymptotically bounded, the limit of the right-hand side of the inequality in (4.43) is 0. Thus, since $F_2 > 0$, we have $\rho_{i,j}^{(2)} \to \mathcal{S}$.

Also, for each pair of relays $(\mathbf{y}_i,\mathbf{y}_j)$ with $i \neq j$, define

$$\begin{aligned}\rho_{i,j}^{(3)} &= \left\{\mathbf{x}\in\mathcal{S} : \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln\mathrm{SNR}^{(i,\mathrm{Rx})} > \mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln\mathrm{SNR}^{(\mathbf{x},j)}\right\}\\ &= \left\{\mathbf{x}\in\mathcal{S} : \frac{2E_iF_2}{N_0\left(C+\|\mathbf{y}_i\|^2\right)} - \ln\alpha_i + \ln\left(\frac{2E_iF_2}{N_0\left(C+\|\mathbf{y}_i\|^2\right)}\right)\right.\\ &\qquad\left. > \frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C+\|\mathbf{x}-\mathbf{y}_j\|^2\right)} - \ln\alpha_j + \ln\left(\frac{2E_{\mathrm{Tx}}F_2}{N_0\left(C+\|\mathbf{x}-\mathbf{y}_j\|^2\right)}\right)\right\}.\\ &\qquad\qquad [\text{from (4.10), (4.23)}]\end{aligned}$$

Observing the symmetry between $\rho_{i,j}^{(3)}$ and $\rho_{i,j}^{(2)}$, we have that as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, $\rho_{i,j}^{(3)}$ becomes either empty, all of $\mathcal{S}$, or the exterior of a circle.

Also, for each pair of relays $(\mathbf{y}_i,\mathbf{y}_j)$ with $i \neq j$, define

$$\begin{aligned}\rho_{i,j}^{(4)} &= \left\{\mathbf{x}\in\mathcal{S} : \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln\mathrm{SNR}^{(i,\mathrm{Rx})}\right.\\ &\qquad\left. > \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln\mathrm{SNR}^{(j,\mathrm{Rx})}\right\}\\ &= \left\{\mathbf{x}\in\mathcal{S} : \frac{2E_iF_2}{N_0\left(C+\|\mathbf{y}_i\|^2\right)} - \ln\alpha_i + \ln\left(\frac{2E_iF_2}{N_0\left(C+\|\mathbf{y}_i\|^2\right)}\right)\right.\end{aligned}$$

$$> \frac{2E_j F_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)} - \ln \alpha_j + \ln\left(\frac{2E_j F_2}{N_0\left(C + \|\mathbf{y}_j\|^2\right)}\right)\Bigg\}.$$

[from (4.10), (4.23)]

Using (4.42), as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, we have $\rho_{i,j}^{(4)} \to \mathcal{S}$ or $\emptyset$.

Then, we have

$$\begin{aligned}
\sigma_{i,j} &= \left\{\mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \left\{\mathbf{x} \in \mathcal{S} : \alpha_i \hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} < \alpha_j \hat{P}_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \Bigg\{\mathbf{x} \in \mathcal{S} : \frac{\alpha_i}{\sqrt{2\pi}} \cdot \max\Bigg(\frac{1}{\sqrt{\mathrm{SNR}^{(\mathbf{x},i)}}} \exp\left\{-\frac{\mathrm{SNR}^{(\mathbf{x},i)}}{2}\right\}, \\
&\qquad \frac{1}{\sqrt{\mathrm{SNR}^{(i,\mathrm{Rx})}}} \exp\left\{-\frac{\mathrm{SNR}^{(i,\mathrm{Rx})}}{2}\right\}\Bigg) \\
&\qquad < \frac{\alpha_j}{\sqrt{2\pi}} \cdot \max\Bigg(\frac{1}{\sqrt{\mathrm{SNR}^{(\mathbf{x},j)}}} \exp\left\{-\frac{\mathrm{SNR}^{(\mathbf{x},j)}}{2}\right\}, \\
&\qquad \frac{1}{\sqrt{\mathrm{SNR}^{(j,\mathrm{Rx})}}} \exp\left\{-\frac{\mathrm{SNR}^{(j,\mathrm{Rx})}}{2}\right\}\Bigg)\Bigg\} \\
&\qquad\qquad \text{[from (4.41)]} \\
&= \Bigg\{\mathbf{x} \in \mathcal{S} : \max\Bigg(\exp\left\{-\frac{\mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln SNR^{(\mathbf{x},i)}}{2}\right\}, \\
&\qquad \exp\left\{-\frac{\mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln SNR^{(i,\mathrm{Rx})}}{2}\right\}\Bigg) \\
&\qquad < \max\Bigg(\exp\left\{-\frac{\mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln SNR^{(\mathbf{x},j)}}{2}\right\}, \\
&\qquad \exp\left\{-\frac{\mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln SNR^{(j,\mathrm{Rx})}}{2}\right\}\Bigg)\Bigg\} \\
&= \Big\{\mathbf{x} \in \mathcal{S} : \min\Big(\mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)}, \\
&\qquad \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(i,\mathrm{Rx})}\Big) \\
&\qquad > \min\Big(\mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)}, \\
&\qquad \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})}\Big)\Big\}
\end{aligned}$$

[for $E_{\mathrm{Tx}}/N_0, E_i/N_0, E_j/N_0$ sufficiently large]

$$
\begin{aligned}
&= \Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)}, \\
&\qquad > \min\Big(\mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)}, \\
&\qquad\qquad \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})}\Big)\Big\} \\
&\quad \cap \Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(i,\mathrm{Rx})} \\
&\qquad > \min\Big(\mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)}, \\
&\qquad\qquad \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})}\Big)\Big\} \\
&= \Big(\Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} > \mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)}\Big\} \\
&\quad \cup \Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(\mathbf{x},i)} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(\mathbf{x},i)} \\
&\qquad > \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})}\Big\}\Big) \\
&\quad \cap \Big(\Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(i,\mathrm{Rx})} \\
&\qquad > \mathrm{SNR}^{(\mathbf{x},j)} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(\mathbf{x},j)}\Big\} \\
&\quad \cup \Big\{\mathbf{x} \in \mathcal{S} : \mathrm{SNR}^{(i,\mathrm{Rx})} - 2\ln\alpha_i + \ln \mathrm{SNR}^{(i,\mathrm{Rx})} \\
&\qquad > \mathrm{SNR}^{(j,\mathrm{Rx})} - 2\ln\alpha_j + \ln \mathrm{SNR}^{(j,\mathrm{Rx})}\Big\}\Big) \\
&= \left(\rho_{i,j}^{(1)} \cup \rho_{i,j}^{(2)}\right) \cap \left(\rho_{i,j}^{(3)} \cup \rho_{i,j}^{(4)}\right).
\end{aligned}
$$

Thus, combining the asymptotic results for $\rho_{i,j}^{(1)}, \rho_{i,j}^{(2)}, \rho_{i,j}^{(3)}$, and $\rho_{i,j}^{(4)}$, as $E_{\mathrm{Tx}}/N_0 \to \infty$, $E_i/N_0 \to \infty$, and $E_j/N_0 \to \infty$, the internal boundary of $\sigma_{i,j}$ consists of circular arcs and line segments. Applying (4.22) completes the proof. ∎

Extended Proof of Lemma 4.4.5. For the upper bound, we have

$$
\begin{aligned}
U\left(\frac{3}{2}, 2, z\right) &= \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_0^\infty e^{-zt} t^{1/2} (1+t)^{-1/2}\, dt \\
&= \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_0^\infty \sqrt{\frac{t}{1+t}} e^{-zt} dt \\
&\le \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_0^\infty e^{-zt} dt \\
&= \frac{1}{\Gamma\left(\frac{3}{2}\right)} \left[-\frac{e^{-zt}}{z}\right]\Bigg|_{t=0}^{\infty} \\
&= \frac{1}{z\Gamma\left(\frac{3}{2}\right)}.
\end{aligned}
$$

For the lower bound, we note that

$$\begin{aligned}
&0 < z < 1 \\
&\Rightarrow 0 < z < 4 \\
&\Rightarrow 0 < \sqrt{z} < 2 \\
&\Rightarrow 0 > -\sqrt{z} > -2 \\
&\Rightarrow 1 > 1 - \sqrt{z} > -1 \\
&\Rightarrow \left(1 - \sqrt{z}\right)^2 < 1 \\
&\Rightarrow 1 - \left(1 - \sqrt{z}\right)^2 > 0
\end{aligned} \tag{4.44}$$

and that

$$\begin{aligned}
&t \geq \frac{(1-\sqrt{z})^2}{\sqrt{z}\,(2-\sqrt{z})} \\
&\Rightarrow t \geq \frac{(1-\sqrt{z})^2}{(1-(1-\sqrt{z}))\,(1+(1-\sqrt{z}))} \\
&\Rightarrow t \geq \frac{(1-\sqrt{z})^2}{1-(1-\sqrt{z})^2} \\
&\Rightarrow t\left(1-(1-\sqrt{z})^2\right) \geq \left(1-\sqrt{z}\right)^2 && \text{[from (4.44)]} \\
&\Rightarrow t \geq \left(1-\sqrt{z}\right)^2 + t\left(1-\sqrt{z}\right)^2 \\
&\Rightarrow \frac{t}{1+t} \geq \left(1-\sqrt{z}\right)^2 \\
&\Rightarrow \sqrt{\frac{t}{1+t}} \geq 1-\sqrt{z}. && [\text{since } 0 < z < 1]
\end{aligned} \tag{4.45}$$

Then, we have

$$\begin{aligned}
U\left(\frac{3}{2}, 2, z\right) &= \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_0^\infty \sqrt{\frac{t}{1+t}} e^{-zt} dt \\
&\geq \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_{\frac{(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}}^\infty \sqrt{\frac{t}{1+t}} e^{-zt} dt && [\text{since } 0 < z < 1] \\
&\geq \frac{1}{\Gamma\left(\frac{3}{2}\right)} \int_{\frac{(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}}^\infty (1-\sqrt{z}) e^{-zt} dt && \text{[from (4.45)]} \\
&= \frac{1-\sqrt{z}}{\Gamma\left(\frac{3}{2}\right)} \left[-\frac{e^{-zt}}{z}\right]\Bigg|_{t=\frac{(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}}^\infty \\
&= \frac{1-\sqrt{z}}{z\Gamma\left(\frac{3}{2}\right)} \exp\left\{-\frac{z(1-\sqrt{z})^2}{\sqrt{z}(2-\sqrt{z})}\right\}
\end{aligned}$$

$$= \frac{1}{z\Gamma\left(\frac{3}{2}\right)}\left(1-\sqrt{z}\right)\exp\left\{-\frac{\sqrt{z}(1-\sqrt{z})^2}{2-\sqrt{z}}.\right\}$$

■

Extended Proof of Theorem 4.4.6. As an approximation to $P_e^{(\mathbf{x},i,\mathrm{Rx})}$ given in (4.15), define

$$\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} = \frac{1}{2} - \left(\frac{D_i\sqrt{\pi}N_0/E_{\mathrm{Tx}}}{8\sigma\left(\sigma^2+B_iN_0/E_{\mathrm{Tx}}\right)^{3/2}}\right)\left(\frac{2\sigma^2\left(\sigma^2+B_iN_0/E_{\mathrm{Tx}}\right)}{\Gamma(3/2)\cdot D_iN_0/E_{\mathrm{Tx}}}\right) \quad (4.46)$$

$$= \frac{1}{2} - \frac{1}{2}\left(1+\frac{1}{2\sigma^2 L_{\mathbf{x},i}E_{\mathrm{Tx}}/N_0}\right)^{-1/2}. \quad \text{[from (4.14)]} \quad (4.47)$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \frac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Using Lemma 4.4.5, it can be shown that

$$\lim_{E_{\mathrm{Tx}}/N_0\to\infty} \alpha_i = 1. \quad (4.48)$$

Specifically, let

$$z = \frac{D_iN_0/E_{\mathrm{Tx}}}{2\sigma^2\left(\sigma^2+B_iN_0/E_{\mathrm{Tx}}\right)} \quad (4.49)$$

and note that $z\to 0$ as $E_{\mathrm{Tx}}/N_0\to\infty$. Then, we have

$$\lim_{E_{\mathrm{Tx}}/N_0\to\infty}\frac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}} = \lim_{E_{\mathrm{Tx}}/N_0\to\infty}\frac{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\cdot U\left(\frac{3}{2},2,z\right)}{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\left(\frac{1}{z\Gamma(3/2)}\right)}$$

[from (4.15), (4.46), (4.49)]

$$\geq \lim_{E_{\mathrm{Tx}}/N_0\to\infty}\frac{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\left(\frac{1}{z\Gamma(3/2)}\right)}{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\left(\frac{1}{z\Gamma(3/2)}\right)} \quad \text{[from Lemma 4.4.5]}$$

$$= 1.$$

We also have

$$\lim_{E_{\mathrm{Tx}}/N_0\to\infty}\frac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$$

$$= \lim_{E_{\mathrm{Tx}}/N_0\to\infty}\frac{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\cdot U\left(\frac{3}{2},2,z\right)}{\frac{1}{2}-\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\left(\frac{1}{z\Gamma(3/2)}\right)} \quad \text{[from (4.15), (4.46), (4.49)]}$$

$$
\begin{aligned}
&\le \lim_{E_{\mathrm{Tx}}/N_0\to\infty} \frac{\frac{1}{2} - \left(\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\right)\left(\frac{1}{z\Gamma\left(\frac{3}{2}\right)}\right)(1-\sqrt{z})\exp\left\{-\frac{\sqrt{z}\left(1-\sqrt{z}\right)^2}{2-\sqrt{z}}\right\}}{\frac{1}{2} - \left(\frac{z\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\right)\left(\frac{1}{z\Gamma(3/2)}\right)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[from Lemma 4.4.5]} \\
&= \lim_{E_{\mathrm{Tx}}/N_0\to\infty} \frac{\frac{1}{2} - \left(\frac{\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\right)\left(\frac{1}{\Gamma\left(\frac{3}{2}\right)}\right)(1-\sqrt{z})\exp\left\{-\frac{\sqrt{z}\left(1-\sqrt{z}\right)^2}{2-\sqrt{z}}\right\}}{\frac{1}{2} - \left(\frac{\sigma\sqrt{\pi}}{4\sqrt{\sigma^2+B_iN_0/E_{\mathrm{Tx}}}}\right)\left(\frac{1}{\Gamma(3/2)}\right)} \\
&= \frac{\frac{1}{2} - \frac{\sigma\sqrt{\pi}}{4\sqrt{\sigma^2}}}{\frac{1}{2} - \frac{\sigma\sqrt{\pi}}{4\sqrt{\sigma^2}}} \\
&= 1.
\end{aligned}
$$

Let

$$
Z_k = \frac{1}{2\sigma^2 L_{\mathbf{x},k}}; \qquad g_k\left(\frac{N_0}{E_{\mathrm{Tx}}}\right) = \sqrt{1+\frac{Z_kN_0}{E_{\mathrm{Tx}}}} - 1 = \left(\frac{Z_k}{2}\right)\frac{N_0}{E_{\mathrm{Tx}}} + \mathcal{O}\left(\left(\frac{N_0}{E_{\mathrm{Tx}}}\right)^2\right) \tag{4.50}
$$

where the second equality in the expression for g_k is obtained using a Taylor series. Then,

$$
\begin{aligned}
\sigma_{i,j} &= \left\{\mathbf{x}\in\mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \left\{\mathbf{x}\in\mathcal{S} : \alpha_i\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} < \alpha_j\hat{P}_e^{(\mathbf{x},j,\mathrm{Rx})}\right\} \\
&= \Bigg\{\mathbf{x}\in\mathcal{S} : \alpha_i\left(\frac{1}{2} - \frac{1}{2}\left(1+\frac{1}{2\sigma^2L_{\mathbf{x},i}E_{\mathrm{Tx}}/N_0}\right)^{-1/2}\right) \\
&\qquad\qquad < \alpha_j\left(\frac{1}{2} - \frac{1}{2}\left(1+\frac{1}{2\sigma^2L_{\mathbf{x},j}E_{\mathrm{Tx}}/N_0}\right)^{-1/2}\right)\Bigg\} \qquad \text{[from (4.47)]} \\
&= \left\{\mathbf{x}\in\mathcal{S} : \alpha_i\left(1-\left(1+\frac{Z_iN_0}{E_{\mathrm{Tx}}}\right)^{-1/2}\right) < \alpha_j\left(1-\left(1+\frac{Z_jN_0}{E_{\mathrm{Tx}}}\right)^{-1/2}\right)\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[from (4.50)]} \\
&= \left\{\mathbf{x}\in\mathcal{S} : \alpha_i\left(1-\frac{1}{\sqrt{1+\frac{Z_iN_0}{E_{\mathrm{Tx}}}}}\right) < \alpha_j\left(1-\frac{1}{\sqrt{1+\frac{Z_jN_0}{E_{\mathrm{Tx}}}}}\right)\right\} \\
&= \left\{\mathbf{x}\in\mathcal{S} : \alpha_i\left(\frac{\sqrt{1+\frac{Z_iN_0}{E_{\mathrm{Tx}}}}-1}{\sqrt{1+\frac{Z_iN_0}{E_{\mathrm{Tx}}}}}\right) < \alpha_j\left(\frac{\sqrt{1+\frac{Z_jN_0}{E_{\mathrm{Tx}}}}-1}{\sqrt{1+\frac{Z_jN_0}{E_{\mathrm{Tx}}}}}\right)\right\}
\end{aligned}
$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i \left(\sqrt{1 + \frac{Z_i N_0}{E_{\text{Tx}}}} - 1 \right) \sqrt{1 + \frac{Z_j N_0}{E_{\text{Tx}}}}}{\alpha_j \left(\sqrt{1 + \frac{Z_j N_0}{E_{\text{Tx}}}} - 1 \right) \sqrt{1 + \frac{Z_i N_0}{E_{\text{Tx}}}}} < 1 \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i \left(\left(\frac{Z_i}{2} \right) \frac{N_0}{E_{\text{Tx}}} + \mathcal{O}\left(\left(\frac{N_0}{E_{\text{Tx}}} \right)^2 \right) \right) \sqrt{1 + \frac{Z_j N_0}{E_{\text{Tx}}}}}{\alpha_j \left(\left(\frac{Z_j}{2} \right) \frac{N_0}{E_{\text{Tx}}} + \mathcal{O}\left(\left(\frac{N_0}{E_{\text{Tx}}} \right)^2 \right) \right) \sqrt{1 + \frac{Z_i N_0}{E_{\text{Tx}}}}} < 1 \right\}$$

[from (4.50)]

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i \left(\frac{Z_i}{2} + \mathcal{O}\left(\frac{N_0}{E_{\text{Tx}}} \right) \right) \sqrt{1 + \frac{Z_j N_0}{E_{\text{Tx}}}}}{\alpha_j \left(\frac{Z_j}{2} + \mathcal{O}\left(\frac{N_0}{E_{\text{Tx}}} \right) \right) \sqrt{1 + \frac{Z_i N_0}{E_{\text{Tx}}}}} < 1 \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i}{\alpha_j} \cdot \frac{\frac{1}{4\sigma^2 L_{\mathbf{x},i}} + \mathcal{O}\left(\frac{N_0}{E_{\text{Tx}}} \right)}{\frac{1}{4\sigma^2 L_{\mathbf{x},j}} + \mathcal{O}\left(\frac{N_0}{E_{\text{Tx}}} \right)} \cdot \sqrt{\frac{1 + \frac{N_0/E_{\text{Tx}}}{2\sigma^2 L_{\mathbf{x},j}}}{1 + \frac{N_0/E_{\text{Tx}}}{2\sigma^2 L_{\mathbf{x},i}}}} < 1 \right\}. \qquad \text{[from (4.50)]}$$

Since $\mathcal{S}$ is bounded, we have, for $E_{\text{Tx}}/N_0 \to \infty$, that

$$\sigma_{i,j} \to \left\{ \mathbf{x} \in \mathcal{S} : \frac{\left(\frac{1}{4\sigma^2 L_{\mathbf{x},i}} \right)}{\left(\frac{1}{4\sigma^2 L_{\mathbf{x},j}} \right)} < 1 \right\} \qquad \text{[from (4.35), (4.48)]}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{1}{4\sigma^2 L_{\mathbf{x},i}} < \frac{1}{4\sigma^2 L_{\mathbf{x},j}} \right\}$$

$$= \{ \mathbf{x} \in \mathcal{S} : L_{\mathbf{x},j} < L_{\mathbf{x},i} \}$$

$$= \{ \mathbf{x} \in \mathcal{S} : \|\mathbf{x} - \mathbf{y}_j\| > \|\mathbf{x} - \mathbf{y}_i\| \} . \qquad \text{[from (4.35), (4.48), (4.23)]}$$

Thus, for $E_{\text{Tx}}/N_0 \to \infty$, the internal boundary of $\sigma_{i,j}$ becomes the line equidistant from $\mathbf{y}_i$ and $\mathbf{y}_j$. Applying (4.22) completes the proof. ■

Extended Proof of Lemma 4.4.7. Using a Taylor series, we have

$$(1+\epsilon)^{1/2} = 1 + \frac{1}{2}\epsilon + \sum_{n=2}^{\infty} \frac{(-1)^{n-1}(2n-2)!}{2^{2n-1}(n-1)!n!}\epsilon^n = 1 + \frac{1}{2}\epsilon - \frac{1}{8}\epsilon^2 + \dots \quad \text{(for } |\epsilon| < 1\text{)}$$

We have

$$1 + \frac{1}{2}\epsilon - \frac{1}{8}\epsilon^2 \le (1+\epsilon)^{1/2} \le 1 + \frac{1}{2}\epsilon \qquad \text{[from a Taylor series]}$$

$$\Rightarrow \frac{1}{1 + \frac{1}{2}\epsilon} \le (1+\epsilon)^{-1/2} \le \frac{1}{1 + \frac{1}{2}\epsilon - \frac{1}{8}\epsilon^2}$$

$$\Rightarrow \frac{1-\left(\frac{1}{1+\frac{1}{x}-\frac{1}{2x^2}}\right)\left(\frac{1}{1+\frac{1}{y}-\frac{1}{2y^2}}\right)}{\frac{1}{x}+\frac{1}{y}} \le \frac{1-\left(1+\frac{2}{x}\right)^{-1/2}\left(1+\frac{2}{y}\right)^{-1/2}}{\frac{1}{x}+\frac{1}{y}} \le \frac{1-\left(\frac{1}{1+\frac{1}{x}}\right)\left(\frac{1}{1+\frac{1}{y}}\right)}{\frac{1}{x}+\frac{1}{y}}$$

$$\Rightarrow \left(\frac{xy}{x+y}\right)\left(1-\left(\frac{x^2}{x^2+x-\frac{1}{2}}\right)\left(\frac{y^2}{y^2+y-\frac{1}{2}}\right)\right) \le L_{x,y}$$
$$\le \left(\frac{xy}{x+y}\right)\left(1-\left(\frac{x}{x+1}\right)\left(\frac{y}{y+1}\right)\right)$$
$$\Rightarrow \left(\frac{xy}{x+y}\right)\frac{\left(x-\frac{1}{2}\right)\left(y^2+y-\frac{1}{2}\right)+x^2\left(y-\frac{1}{2}\right)}{\left(x^2+x-\frac{1}{2}\right)\left(y^2+y-\frac{1}{2}\right)} \le L_{x,y}$$
$$\le \left(\frac{x+y+1}{x+y}\right)\left(\frac{x}{x+1}\right)\left(\frac{y}{y+1}\right).$$

Now, note that for x, y sufficiently large

$$\frac{\left(x-\frac{1}{2}\right)\left(y^2+y-\frac{1}{2}\right)+x^2\left(y-\frac{1}{2}\right)}{\left(x^2+x-\frac{1}{2}\right)\left(y^2+y-\frac{1}{2}\right)} \ge \frac{(x-1)\left(y^2+y-2\right)+\left(x^2-1\right)(y-1)}{\left(x^2+x\right)\left(y^2+y\right)}$$
$$= \frac{(x-1)(y-1)(x+y+3)}{xy(x+1)(y+1)}.$$

Thus, we have

$$\left(\frac{x-1}{x+1}\right)\left(\frac{y-1}{y+1}\right)\left(\frac{x+y+3}{x+y}\right) \le L_{x,y} \le \left(\frac{x+y+1}{x+y}\right)\left(\frac{x}{x+1}\right)\left(\frac{y}{y+1}\right).$$

[for x, y sufficiently large]

Now taking the limit as $x \to \infty$ and $y \to \infty$ (in any manner) gives $L_{x,y} \to 1$. ■

Extended Proof of Theorem 4.4.8. As an approximation to $P_e^{(\mathbf{x},i,\mathrm{Rx})}$ given in (4.18), define

$$\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} = \frac{1/2}{\mathrm{SNR}^{(\mathbf{x},i)}} + \frac{1/2}{\mathrm{SNR}^{(i,\mathrm{Rx})}}. \tag{4.51}$$

For any relay $\mathbf{y}_i$, let $\alpha_i = \dfrac{P_e^{(\mathbf{x},i,\mathrm{Rx})}}{\hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})}}$. Using Lemma 4.4.7, it can be shown that

$$\lim_{\substack{E_{\mathrm{Tx}}/N_0 \to \infty, \\ E_i/N_0 \to \infty}} \alpha_i = 1. \tag{4.52}$$

Then, we have

$$\sigma_{i,j} = \left\{\mathbf{x} \in \mathcal{S} : P_e^{(\mathbf{x},i,\mathrm{Rx})} < P_e^{(\mathbf{x},j,\mathrm{Rx})}\right\}$$
$$= \left\{\mathbf{x} \in \mathcal{S} : \alpha_i \hat{P}_e^{(\mathbf{x},i,\mathrm{Rx})} < \alpha_j \hat{P}_e^{(\mathbf{x},j,\mathrm{Rx})}\right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{\alpha_i}{\mathrm{SNR}^{(\mathbf{x},i)}} + \frac{\alpha_i}{\mathrm{SNR}^{(i,\mathrm{Rx})}} < \frac{\alpha_j}{\mathrm{SNR}^{(\mathbf{x},j)}} + \frac{\alpha_j}{\mathrm{SNR}^{(j,\mathrm{Rx})}} \right\} \quad \text{[from (4.51)]}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \alpha_i \cdot \frac{N_0 \|\mathbf{x} - \mathbf{y}_i\|^2}{E_{\mathrm{Tx}}} + \alpha_i \cdot \frac{N_0 \left(C + \|\mathbf{y}_i\|^2\right)}{E_i} \right.$$
$$\left. < \alpha_j \cdot \frac{N_0 \|\mathbf{x} - \mathbf{y}_j\|^2}{E_{\mathrm{Tx}}} + \alpha_j \cdot \frac{N_0 \left(C + \|\mathbf{y}_j\|^2\right)}{E_j} \right\}$$
$$\text{[from (4.8), (4.23)]}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \alpha_i \|\mathbf{x} - \mathbf{y}_i\|^2 + \alpha_i \left(C + \|\mathbf{y}_i\|^2\right) \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0} \right.$$
$$\left. < \alpha_j \|\mathbf{x} - \mathbf{y}_j\|^2 + \alpha_j \left(C + \|\mathbf{y}_j\|^2\right) \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \alpha_i \left(\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}_i \rangle + \|\mathbf{y}_i\|^2\right) - \alpha_j \left(\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}_j \rangle + \|\mathbf{y}_j\|^2\right) \right.$$
$$\left. < \alpha_j \left(C + \|\mathbf{y}_j\|^2\right) \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} - \alpha_i \left(C + \|\mathbf{y}_i\|^2\right) \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0} \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : (\alpha_i - \alpha_j) \|\mathbf{x}\|^2 + 2\alpha_j \langle \mathbf{x}, \mathbf{y}_j \rangle - 2\alpha_i \langle \mathbf{x}, \mathbf{y}_i \rangle + \alpha_i \|\mathbf{y}_i\|^2 - \alpha_j \|\mathbf{y}_j\|^2 \right.$$
$$\left. < \alpha_j C + \|\mathbf{y}_j\|^2 \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} - \alpha_i C + \|\mathbf{y}_i\|^2 \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0} \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \frac{(\alpha_i - \alpha_j) \|\mathbf{x}\|^2}{2} + \langle \mathbf{x}, \alpha_j \mathbf{y}_j \rangle - \langle \mathbf{x}, \alpha_i \mathbf{y}_i \rangle + \frac{\alpha_i \|\mathbf{y}_i\|^2}{2} - \frac{\alpha_j \|\mathbf{y}_j\|^2}{2} \right.$$
$$\left. < \frac{\alpha_j \left(C + \|\mathbf{y}_j\|^2\right)}{2} \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} - \frac{\alpha_i \left(C + \|\mathbf{y}_i\|^2\right)}{2} \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0} \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{S} : \langle \mathbf{x}, \alpha_j \mathbf{y}_j - \alpha_i \mathbf{y}_i \rangle \right.$$
$$< \frac{\alpha_j \left(C + \|\mathbf{y}_j\|^2\right)}{2} \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_j/N_0} - \frac{\alpha_i \left(C + \|\mathbf{y}_i\|^2\right)}{2} \cdot \frac{E_{\mathrm{Tx}}/N_0}{E_i/N_0}$$
$$\left. + \frac{(\alpha_j - \alpha_i) \|\mathbf{x}\|^2}{2} + \frac{\alpha_j \|\mathbf{y}_j\|^2}{2} - \frac{\alpha_i \|\mathbf{y}_i\|^2}{2} \right\}.$$

Now, for any relay $\mathbf{y}_k$, let

$$G_k = \lim_{\substack{E_{\mathrm{Tx}}/N_0 \to \infty, \\ E_k/N_0 \to \infty}} \frac{E_k/N_0}{E_{\mathrm{Tx}}/N_0}.$$

Using (4.52), Table 4.1 considers the cases of G_i and G_j being zero, infinite, or finite non-zero; for all possible combinations, the internal boundary of $\sigma_{i,j}$ is linear. Applying (4.22) completes the proof. ■

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, DC: National Bureau of Standards, 1966.

[2] J. Balam and J. D. Gibson, "Adaptive event coverage using high power mobiles over a sensor field," in *Proc. IEEE 60th Veh. Technol. Conf.*, Los Angeles, CA, Sept. 2004, pp. 4611 – 4615.

[3] D. Chen and J. N. Laneman, "Modulation and demodulation for cooperative diversity in wireless systems," *IEEE Trans. Wireless Commun.*, vol. 5, no. 7, pp. 1785 – 1794, 2006.

[4] Y. Chen, Z. Wang, and J. Liang, "Automatic dynamic flocking in mobile actuator sensor networks by central Voronoi tessellations," in *Proc. IEEE Int. Conf. on Mechatronics and Automation*, Niagara Falls, Canada, Jul. 2005, pp. 1630 – 1635.

[5] W. Cho and L. Yang, "Energy and location optimization for relay networks with differential modulation," in *25th Army Science Conf.*, Orlando, FL, Nov. 2006.

[6] C. Comaniciu and H. V. Poor, "On the capacity of mobile ad hoc networks with delay constraints," *IEEE Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2061 – 2071, 2006.

[7] J. Cortés, S. Martínez, T. Karataş, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Trans. Robot. Autom.*, vol. 20, no. 2, pp. 243 – 255, 2004.

[8] S. C. Ergen and P. Varaiya, "Optimal placement of relay nodes for energy efficiency in sensor networks," in *Proc. IEEE Int. Conf. on Communications (ICC)*, Istanbul, Turkey, Jun. 2006, pp. 3473 – 3479.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic Publishers, 1992.

[10] M. O. Hasna and M.-S. Alouini, "End-to-end performance of transmission systems with relays over Rayleigh-fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1126 – 1131, 2003.

[11] Y. T. Hou, Y. Shi, H. D. Sherali, and S. F. Midkiff, "On energy provisioning and relay node placement for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2579 – 2590, 2005.

[12] A. Iranli, M. Maleki, and M. Pedram, "Energy efficient strategies for deployment of a two-level wireless sensor network," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, San Diego, CA, Aug. 2005, pp. 233 – 238.

[13] I. Koutsopoulos, S. Toumpis, and L. Tassiulas, "On the relation between source and channel coding and sensor network deployment," in *Proc. Int. Workshop on Wireless Ad-hoc Networks (IWWAN)*, London, England, May 23-26, 2005.

[14] X. Liu and P. Mohapatra, "On the deployment of wireless sensor nodes," in *Proc. 3rd Int. Workshop on Measurement, Modeling, and Performance Analysis of Wireless Sensor Networks (SenMetrics)*, San Diego, CA, Jul. 2005, pp. 78 – 85.

[15] Y. Mao and M. Wi, "Coordinated sensor deployment for improving secure communication and sensing coverage," in *Proc. 3rd ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN)*, Alexandria, VA, November 7, 2005.

[16] L. Ong and M. Motani, "On the capacity of the single source multiple relay single destination mesh network," *Ad Hoc Networks*, vol. 5, no. 6, pp. 786 – 800, 2007.

[17] J. Pan, L. Cai, Y. T. Hou, Y. Shi, and S. X. Shen, "Optimal base-station locations in two-tiered wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 4, no. 5, pp. 458 – 473, 2005.

[18] J. G. Proakis, *Digital Communications*, 4th edition, New York, NY: McGraw-Hill, 2001.

[19] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 1996.

[20] A. K. Sadek, Z. Han, and K. J. Liu, "An efficient cooperation protocol to extend coverage area in cellular networks," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, Las Vegas, NV, Apr. 2006, pp. 1687 – 1692.

[21] A. So and B. Liang, "Exploiting spatial diversity in rate adaptive WLANs with relay infrastructure," in *Proc. IEEE Global Telecommunications Conf. (GLOBECOM)*, St. Louis, MO, Nov. 2005.

[22] J. Suomela, "Approximating relay placement in sensor networks," in *Proc. 3rd ACM Int. Workshop on Evaluation of Wireless Ad Hoc, Sensor and Ubiquitous Networks (PE_WASUN)*, Terromolinos, Spain, Oct. 2006, pp. 145 – 148.

[23] J. Tan, O. M. Lozano, N. Xi, and W. Sheng, "Multiple vehicle systems for sensor network area coverage," in *Proc. 5th World Conf. on Intelligent Control and Automation*, Hangzhou, China, Jun. 2004, pp. 4666 – 4670.

[24] H. L. van Trees, *Detection, Estimation and Modulation Theory, Part 1*. New York, NY: John Wiley & Sons, 1968.

[25] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York, NY: John Wiley & Sons, 1965.

Chapter 5

Conclusion

This thesis considered three communications problems in the areas of network coding and wireless sensor networks. The main contributions are now summarized and possible directions for future research are discussed.

Chapter 2 formally defined the routing capacity of a network and showed that it is rational, achievable, and computable. While it is known that the (general) coding capacity of a network it not necessarily achievable, it would be interesting to study these properties for the general coding capacity as well as for the linear coding capacity. In particular, the existence of a general algorithm for finding the coding capacity of network would be significant. Similarly, determining a more efficient algorithm for finding the routing capacity than that presented in this thesis would be of practical importance. Relations between the routing, linear, and general coding capacity of a network (such as when one is strictly larger than another) would also provide theoretical insight into network coding.

Chapter 3 formally defined the uniform and average node-limited coding capacities of a network and showed that every non-negative, monotonically non-decreasing, eventually-constant, rational-valued function on the integers is the node-limited capacity of some network. An immediate method of extending the average coding capacity definition would be to use a weighted sum of coding rates. The weighting coefficients would allow preference to be given to specific source messages. Determining properties of the weighted node-limited capacity would parallel the work in this thesis. It would also be of theoretical interest to determine whether or not the node-limited coding capacity of a network can have some irrational and some rational values, or some achievable and some unachievable values.

Chapter 4 gave an algorithm that determines relay positions and sensor-relay assignments in wireless sensor networks. Communications were modeled using path loss, fading, and additive white Gaussian noise, and the algorithm attempted to minimize the probability of error at the receiver. Analytic expressions, with respect to fixed relay positions, describing the sets of locations in the plane in which sensors are (optimally) assigned to the same relay were given for both amplify-and-forward and decode-and-forward relays protocols, in some instances for the case of high transmission energy per bit. Numerical results showing the output of the algorithm, evaluating its performance, and examining the accuracy of the high power approximations were also presented. To enhance the relay placement algorithm, the system model used for the wireless sensor network could be extended. The inclusion of multi-hop relay paths would provide a more realistic setting. Incorporating more sophisticated diversity combining techniques would also improve the network performance and increase the applicability of the algorithm. Much of the analysis of these this thesis holds for higher order path loss; thus, extending the model to allow the path loss exponent to be a function of distance would more closely approximate real-world situations. Including power constraints and allowing relays to use different gains are also interesting problems. Introducing priorities on the sensor nodes would add more generality to the model. Finally, exploiting correlation between the sensors would be a natural extension and would improve system performance.