ACCEPTED FROM OPEN CALL

Beyond Shannon: The Quest for Fundamental Performance Limits of Wireless Ad Hoc Networks

Andrea Goldsmith, Stanford University
Michelle Effros, Caltech
Ralf Koetter, Technical University of Munich
Muriel Médard, Asu Ozdaglar, and Lizhong Zheng, MIT

ABSTRACT

We describe a new theoretical framework for determining fundamental performance limits of wireless ad hoc networks. The framework expands the traditional definition of Shannon capacity to incorporate notions of delay and outage. Novel tools are described for upper and lower bounding the network performance regions associated with these metrics under a broad range of assumptions about channel and network dynamics, state information, and network topologies. We also develop a flexible and dynamic interface between network applications and the network performance regions to obtain the best end-to-end performance. Our proposed framework for determining performance limits of wireless networks embraces an interdisciplinary approach to this challenging problem that incorporates Shannon Theory along with network theory, combinatorics, optimization, stochastic control, and game theory. Preliminary results of this approach are described and promising future directions of research are outlined.

INTRODUCTION

A wireless ad hoc network is a collection of wireless nodes that self-configure to form a network without the aid of any established infrastructure. When these networks have mobile nodes, as shown in Fig. 1, they are called mobile ad hoc networks (MANETs). Without an inherent infrastructure, the nodes perform the required network control and management through dynamic control algorithms. Multihop routing, whereby intermediate nodes relay data toward its final destination, is typically used to increase network performance and throughput as well as the distances over which network source and destination nodes can communicate. Network nodes typically communicate through bi-directional communication links, and feedback channels (either separate or piggybacked on the direct links) may also be available.

Wireless ad hoc networks are highly appealing for many reasons. They can be rapidly deployed and reconfigured. They can be tailored to specific applications. They are also highly robust due to their distributed nature, node redundancy, and the lack of single points of failure. Robustness is especially important in military applications for which the first wireless ad hoc network protocols were developed. However, despite much research activity over the last several decades on wireless communications in general, and on wireless ad hoc networks in particular, there remain many significant technical challenges in the design and performance optimization of these networks. In particular, the Shannon capacity region of wireless ad hoc networks — the region defining the maximum rates achievable between all node pairs — has remained an open problem for decades, even for the most simple network topologies. The fundamental performance of wireless networks for metrics other than capacity is also poorly understood. These fundamental performance bounds could provide insight to improve network design and performance, as well as an upper bound against which to compare the performance of existing protocols, as Shannon capacity has done for point-to-point and multiuser channels.

The mathematical theory of information was born from Claude Shannon's conception of channel capacity in 1948 — the maximum rate that can be achieved over a channel with asymptotically small probability of error. The simple yet elegant mathematics of this brilliant concept, coupled with its revolutionary ideas for coding over noisy channels and bounding their fundamental data rate limits via their mutual information, has inspired generations of theorists and practitioners. Moreover, much insight as well as design breakthroughs in communication systems, such as multiple-input multiple-output (MIMO) techniques, water-filling rate and power adapta-

This work was supported by the DARPA ITMANET program under grant 1105741-1-TFIND.

tion, and low-density parity check (LDPC) codes, resulted from Shannon's results and those that built upon then. Modular system design, in particular the separation of application-layer compression techniques and lower-layer transmission protocols, were also inspired by Shannon's results on the optimality of source and channel code separation.

There has been much progress in finding the Shannon capacity of wireless single-user and multiuser channels. For single-user channels this capacity is a number, the maximum data rate of the channel, while for multiuser channels it is an n-dimensional region defining the maximum rates possible for all n users simultaneously, as shown in Fig. 2 for a three-user channel. The capacity of static single-user channels, multiple access channels (MACs), and broadcast channels (BCs) with noise and multipath is well known. In recent years these results have been extended to include multiple antennas at the transmitter and receiver (MIMO channels) [1], where the additional spatial dimension increases capacity linearly with the number of antennas.

The capacity of time-varying flat-fading single-user, MAC, and BC channels when the transmitter(s) and receiver(s) have perfect information about the channel state is also well known. Without channel state information at the transmitter, capacity is often limited by the worst-case channel conditions, which can be quite bad in wireless channels. To address this issue, the definition of capacity in time-varying channels has been modified to include error in the form of outage, and results for this metric under different fading distributions in the single-user, MAC, and BC have been derived [1]. Capacity of channels with feedback is largely unknown, except for the case of single-user static memoryless channels, where feedback generally doesn't increase capacity, and time-varying finite-state channels, where the structure in the channel variations allows the impact of feedback on capacity to be determined. The lack of capacity results for feedback channels is reflected in the ad hoc use of feedback in wireless system design, which typically consists of channel and network state information as well as acknowledgments of or retransmission requests for transmitted packets. Yet there is no fundamental result that dictates this is the best use of the finite-rate feedback channels built into these systems.

Capacity results are much more limited for ad hoc wireless networks, even for simple static models. In particular, the capacities of the most basic ad hoc networks, the three-node relay channel and the four-node interference channel illustrated in Fig. 3, have remained open problems for decades. There has been significant progress in deriving capacity scaling laws, which characterize how per-node capacity scales in an asymptotically large network [2]. However, these laws provide just one point, the equal-rate point, on the wireless network capacity region; for n users this region has dimension $n \times (n-1)$ since it dictates the maximum data rates between all pairs of users simultaneously. Similarly, interference alignment can achieve the sum capacity rate point (the maximum of the sum of all user rates) in interference networks, but does not achieve the full capacity region [3]. Since the fundamental data rate limits of the wireless networks shown in Fig. 3 have eluded researchers for so long, it is unlikely that we can obtain the Shannon capacity region of more general networks, especially when network dynamics and feedback are incorporated. Separation theorems to guide wireless network protocol designs are also absent due to the lack of capacity results for these networks.

The Shannon capacity of a channel places no restrictions on complexity or delay in transmission or reception. Technology evolution has borne out the appropriateness of the first assumption, as silicon chips today support highly-complex designs with relatively small, cheap, and low-power implementations unimaginable in Shannon's day. In contrast, however, the Shannon capacity assumption of asymptotically large delays is problematic for real-time applications and gives rise to an unconsummated union between network theory and Shannon theory [4], since the former is largely based on analysis of queuing delays. There has been very limited progress and therefore little insight into consummating this union to obtain the fundamental

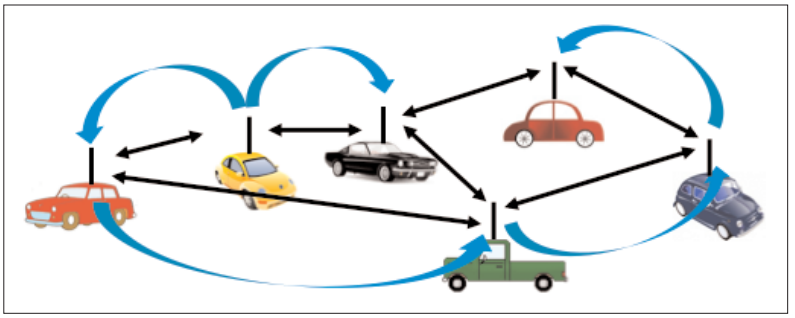

Figure 1. *A mobile ad hoc network (MANET). Black arrows indicate bidirectional communication links, while blue arrows indicate feedback channels (which may be separate or piggybacked on the direct links).*

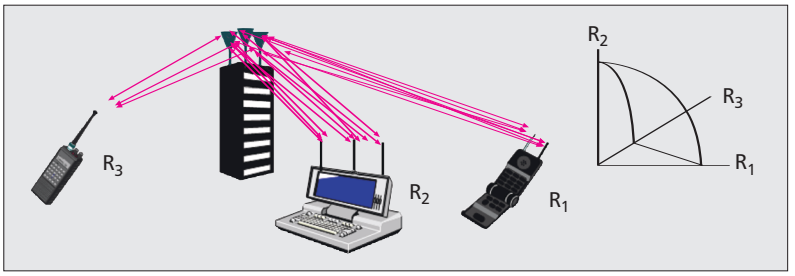

Figure 2. *Capacity region for three users.*

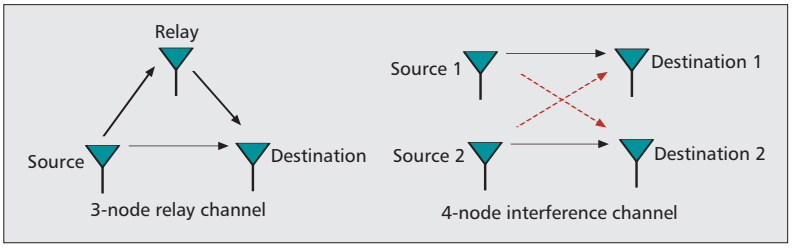

Figure 3. *Simple ad hoc networks for which capacity is unknown.*

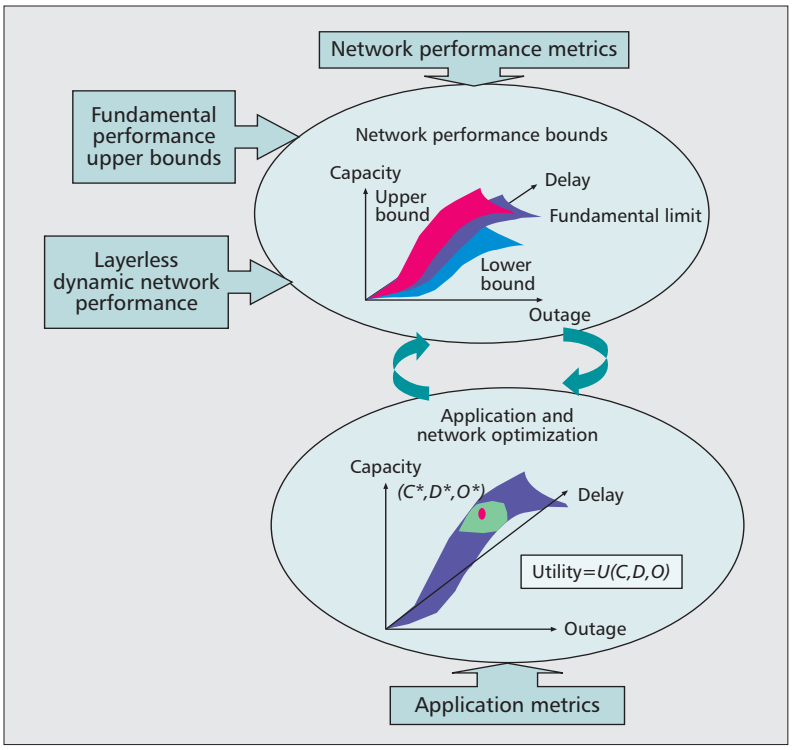

Figure 4. *A framework for obtaining fundamental performance limits of wireless ad hoc networks.*

capacity limits of wireless systems with delay constraints.

The grand research challenge of developing and exploiting a more powerful Information Theory for MANETs initiated the DARPA Information Theory of Mobile Ad Hoc Networks (ITMANET) program. The program's hypothesis is that a better understanding of these fundamental capacity limits will lead to insights and implications for network design and deployment, including an optimal layering of the protocol stack defining the appropriate interface between applications and their underlying networks. This article provides an overview of our research under the ITMANET program toward developing this new theory. In particular, we will provide an overview of the many research challenges in developing this theory, as well as our research framework for breaking the problem into separate yet interconnected research areas to address these challenges. These research areas are both evolutionary and revolutionary with respect to the current state-of-the-art, and in fact the interconnection between the areas is what we believe will provide the theoretical breakthroughs toward this new theory. The principle researchers on the project include the authors of this article[1] as well as Stephen Boyd, Todd Coleman, Ramesh Johari, Sean Meyn, Pierre Moulin, and Devavrat Shah. The team brings an interdisciplinary perspective and rich set of mathematical tools to bear on this complex problem, as described in the next section. Note that a previous paper [5] described a different approach to these challenges: a notion of functional network information theory that provides useful upper bounds on network throughput which are robust to nonideal assumptions, encompass delay, and can be approached by real designs in the foreseeable future. We take a different approach: that design breakthroughs and insight in wireless ad hoc networks require fundamental, rather than functional, upper and lower bounds on performance, but these bounds can only be obtained through an interdisciplinary approach that combines information theory with network theory, optimization, and control. These additional tools are needed to address the complexity of performance bounds for large networks, as well as the random dynamics of traffic and network topology, which are not easily handled by traditional information-theoretic tools.

Fundamental Performance Limits: An Interdisciplinary Approach

The quest for fundamental performance limits of wireless ad hoc networks is a highly complex problem. To better manage its scope and make progress, our research framework breaks the problem into three interconnected research areas — fundamental performance upper bounds, layerless dynamic network performance, and application and network optimization — as illustrated in Fig. 4. The first research area explores new paradigms for upper bounds on wireless ad hoc network performance, while the second explores novel "layerless" networking techniques to lower bound this performance. In this context a layerless network does not impose any predefined protocol structure that might result in suboptimal performance. The rationale for separately exploring upper and lower bounds is that they typically entail different tools and perspectives. Upper bounds exploit abstract mathematical constructs that incorporate the performance of any possible scheme, while lower bounds are simpler in that any concrete scheme provides a lower bound. However, tight upper and lower bounds require insight about each other, and it is at the intersection of these two research areas — when the upper and lower bounds meet — that the fundamental performance limits become known. The last research area develops the interface between the application and network to achieve the best end-to-end performance. Investigation of end-to-end network performance, including optimal routing and scheduling protocols as well as throughput versus delay tradeoffs, typically relies heavily on analytical tools from control and optimization, rather than the information-theoretic and combinatorial tools used in most results to date on fundamental upper and lower performance bounds. This difference in perspectives and tools is largely responsible for the unconsummated union between information theory and network theory [4]. We believe progress can be made in consummating this union by incorporating the coding and relaying strategies developed for layerless dynamic networks that approach the network performance upper bounds into the network models to which we apply our optimization and control tools. Underlying all three of

[1] *Ralf Koetter passed away prior to publication of this article.*

our research areas are the models and metrics for the networks we analyze; which metrics are important in terms of network performance, and which canonical models are both insightful and tractable to analyze with respect to these metrics.

Figure 4 illustrates a performance region where capacity is not the only metric. Indeed, as discussed earlier, delay (its average, maximum, or distribution) is an important design aspect in many applications. In addition, dynamic wireless channels may exhibit improved rates if some outage or error is allowed. Hence the performance region in Fig. 4 shows a hypothetical tradeoff between data rate (capacity), delay, and outage for a given user. But these are not the critical metrics for every system or user. For example, sensor networks with non-rechargeable batteries may wish to optimize energy per bit rather than data rate; communication systems where reliability is tantamount may preclude any outage; and data systems may operate with little constraint on delay. Thus, the figure merely illustrates the performance region for a user where data rate, delay, and outage are the most important metrics. Note that transmit power is not explicit in this performance region but rather is a parameter of the underlying model. Other model parameters might include available bandwidth, number of antennas at each node, and complexity limitations.

Upper and lower network performance bounds dictate design tradeoffs available at the application layer of the network: e.g., a higher data rate might be achieved but at the expense of some delay; or if outage is precluded then data rates will typically suffer. Where to operate on this network performance region depends on the application. In particular, for the region of Fig. 4, the optimal operating point will depend on how sensitive the application is to the outage, delay, and data rate of the network. A video application might operate at a high rate with low delay, but suffer outage when network conditions do not support this level of performance. A voice application might operate at a relatively low data rate with minimal outage and strict delay constraints. Networks might implement multiple tiers of service, where some users have better performance while other users with lower priority have worse performance. All of these network designs correspond to different operating points on the performance region illustrated in Fig. 4. The goal of our research in application and network optimization is to determine this optimal operating point for each user based on the application metrics and the underlying performance region. These three research areas comprise an interdisciplinary approach beyond traditional Shannon Theory that could hold much promise for breakthroughs in establishing the fundamental performance limits of wireless ad hoc networks and the applications that utilize them.

The remainder of this article describes the existing and emerging intellectual tools and methods as well as our preliminary results within this research framework. In particular, in the area of upper bounds we describe the new notion of network equivalence as well as some recent breakthroughs on strong converses. Layerless dynamic networks utilize advances in network coding and optimized feedback. In addition, we describe how traditional notions of capacity for these networks can incorporate outage and delay through the use of hierarchical codes. Finally, advances in network utility maximization, generalized MaxWeight policies for scheduling and routing, and stochastic game theory have provided breakthroughs in application and network optimization as well as interconnections with research in upper and lower network performance bounds. These areas of investigation are illustrated in Fig. 5 and described in more detail in subsequent sections. Note that these subsequent sections will cover a rich and varied tapestry of past and current research. Our goal is not to provide a tutorial review of these topics, but rather to discuss how they fit within our vision for establishing a new research framework to investigate fundamental performance limits of wireless ad hoc networks, and our relevant work in this domain. The limited references we provide are by no means comprehensive but rather serve as a starting point for further reading.

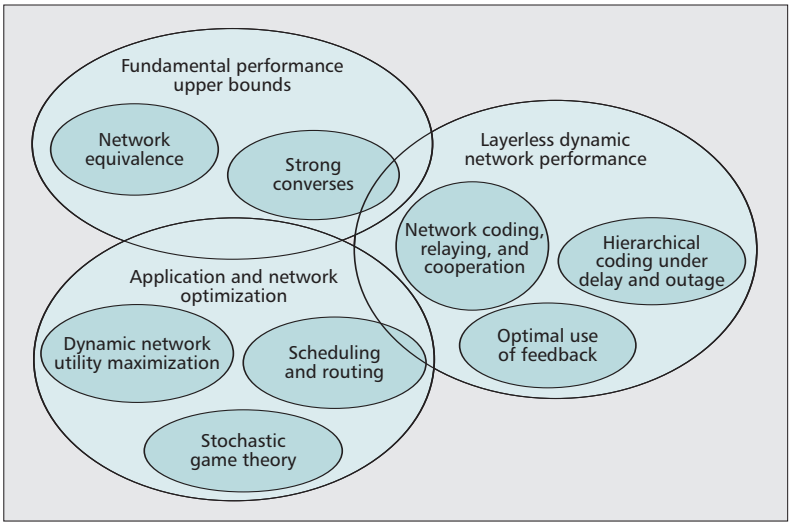

Figure 5. *Existing and emerging analytical tools within our research framework.*

Fundamental Upper Bounds on Network Performance

The usual approach for finding capacity regions is to separately find upper and lower bounds, which may not and often do not coincide. Of the two types of bounds, lower bounds are much easier to come by, since basically any scheme, no matter how simple or naive, yields a lower bound on optimal performance. Upper bounds, which define the best performance possible for any arbitrary scheme, are much harder to obtain since it must be verified that every possible scheme has performance no better than that of the upper bound. Of course, the ultimate goal is to propose a *really good* networking strategy whose performance meets that of the upper bound, thereby yielding the capacity region.

Most capacity upper bounds rely on Fano's inequality and the cut-set bounds derived from

While solving the network coding capacity problem on networks of noiseless links is computationally challenging, complete solutions for small networks and approximations for larger networks both pose reasonable paths for capacity calculation in a wide range of wireless and wireline networks.

it. Cut-set bounds dictate that the rate transmitted across a cut in the network is no greater than what could be transmitted if all nodes on the transmitting side of the cut worked together as a single transmitter node and all nodes on the receiving side of the cut worked together as a single receiver node. Cut-set bounds are commonly used to upper-bound capacity of small networks, such as the canonical configurations of Fig. 3, as well as to obtain scaling laws. Cut-set bounds are broadly applicable to most network topologies and are typically analytically tractable and easy to work with. Unfortunately, though, these bounds do not extend easily to networks with multiple sources and destinations and are known to be loose for most networks of interest. In summary, the traditional tools of Fano's inequality and cut-set bounds have been of limited value in tightly upper bounding capacity of typical wireless ad hoc network configurations or in gleaning much insight into their performance limits. This state of affairs indicates a clear need for new tools and approaches to upper bound wireless ad hoc network capacity. The most promising approaches we have uncovered in our work for obtaining these bounds is using either strong converse proofs or network equivalence.

Strong Converses

Capacity theorems are typically proved using either strong or weak converse theorems. Roughly speaking, the weak converse states that the probability of error (average or maximum) at rates above capacity is bounded away from zero, while for the strong converse this probability approaches one. The weak converse is often obtained using Fano's inequality, which as stated above is the basis for most capacity upper bounds. Our recent work indicates that some open multiuser capacity problems can be solved based on the more stringent strong converse. In particular, in [6] the capacity region for the multiple access channel with channel state information at the transmitters has been derived based on a strong converse with a maximum error criteria. The weak converse yields a much looser rate region that does not coincide with known achievable rates. Moreover, the strong converse provides more insight into the optimal coding strategy, since it aligns with the achievable rate region based on random binning. While these results apply only to multiple access channels, they indicate that strong converses are a promising area to explore in cracking open capacity problems where the upper bound based on a weak converse is loose compared to the best known achievable rate.

Network Equivalence

We have also proposed a *new approach* to capacity upper bounds based on the notion of network equivalence. Specifically, it is shown in [7] that, given a network of noisy, independent, memoryless links, a set of user rate demands can be met on the given network if and only if it can be met on another network where each noisy link is replaced by a noiseless bit pipe of the noisy link capacity. This network equivalence is proved by showing that if we know how to operate a network of noisy links at a given rate point, we can find a way to operate the corresponding network of noiseless links at the same rate point. This approach never answers the question of which rate points are achievable. It only demonstrates the equivalence of the capacity regions of two networks — one with noisy links and the other with noiseless links of the corresponding capacities. This result also proves the optimality of separating the designs of the network and channel coding strategies over point-to-point links; this separation is not unlike Shannon's result on the optimality of source and channel coding separation for a single point-to-point link, though it does not follow from that result.

This equivalence approach for point-to-point channels has also been applied to networks containing memoryless BCs, MACs, interference channels, and more general channels [7]. In this notion of equivalence, for each channel a network of noiseless links is found whose performance bounds the component behavior either from below or from above. The capacity of an arbitrary network comprised of the given components is then bounded by the capacity of the network constructed by replacing each channel by its bounding noiseless model. (Replacing each channel with its lower bounding model yields a lower bound on capacity; replacing each channel by its upper bounding model yields an upper bound on capacity.) Finding the network coding capacities of these deterministic networks gives upper and lower bounds on the network capacity. Comparing the upper and lower bounds to each other yields bounds on their accuracy.

In [8], Avestimehr, Diggavi, and Tse also relate the capacities of wireless networks to those of deterministic networks. Their deterministic models contain both noiseless point-to-point links and deterministic MACs. Their results apply only to networks for which good prior outer bounds on network capacity are already available. As a result, their deterministic model is applied primarily in cases where cut-set bounds are known to be tight. In these cases, they demonstrate achievability results and bound the distance of these rates from prior upper bounds on the capacity region. While their results and techniques are quite different, we believe it to be no coincidence that both approaches reduce a stochastic problem to its combinatoric core.

Combining network equivalence with existing tools for the analysis of network coding capacities provides a general strategy for systematic network capacity analysis. While solving the network coding capacity problem on networks of noiseless links is computationally challenging, complete solutions for small networks and approximations for larger networks both pose reasonable paths for capacity calculation in a wide range of wireless and wireline networks.

Some Open Challenges

While strong converses provide a path forward for some network capacity upper bounds, they become highly complex as the number of nodes in the network grows. Hence, the use of strong or weak converses to tightly upper bound performance of reasonably-sized networks is limited, and perhaps a different set of bounding tools is

needed. Network equivalence is applicable to any size network, but its computational complexity grows with network size. Other new upper bounding tools such as deterministic models are most useful where tight upper bounds on performance already exist, and are also difficult to extend to network models of more than a few nodes. Breakthroughs are needed to find new upper bounding techniques applicable to networks of more than a few nodes with reasonable complexity. In addition, incorporating notions of delay and outage into upper bounds on performance of multihop networks remains another significant challenge. In particular, while results exist that incorporate outage into the capacity of point-to-point, MAC, and BC channels, these results have not yet been extended to more general networks.

Layerless Dynamic Network Performance

One of the biggest challenges in obtaining fundamental performance limits of wireless ad hoc networks is their dynamics. These dynamics occur on different time scales; the wireless channel varies much faster than the network traffic, which in turn varies faster than topological changes as users enter and leave the network. Due to these different time scales of variation, the protocols that govern communication networking, from access controls, routing, and congestion control to the exchange of network and channel state information to the coding, modulation, and MIMO strategies over the physical channels, all must be designed relative to the dynamic requirements at the time scales over which they occur.

The different time scales in wireless networks coupled with the engineering principle of breaking a complex problem into multiple simpler problems has led to a layered architecture for wireless network protocols. Such separation greatly simplifies network designs via a divide-and-conquer procedure. However, recent results show that wireless networks which break this paradigm through a cross-layer design can exhibit better performance and flexibility, although caution must be taken to preserve the benefits of architecture in the protocol stack as well as avoid inadvertent protocol interactions [9]. In order to consider novel transmission strategies and the underlying network performance limits without presupposed notions of protocol layering, we take the network architectural view of *layerless* design.

The theoretical foundation of layered design in digital communications is Claude Shannon's *separation theorem*, where it was shown that separate design of information compression and channel coding does not cause any performance loss, provided that the underlying communication channel is point-to-point and static. However, in wireless networks, the underlying channel capacity can change over time and data can be sent to multiple destinations, challenging Shannon's key assumptions associated with channel capacity and separation. Moreover, the source and channel codes in networks are separated by the network protocol (or "network code"), which governs channel access, routing, and other end-to-end network functions. Surely if the source and channel code designs cannot be optimally separated from each other in wireless channels, neither can be optimally separated from the network protocol stack governing end-to-end transmission that lies between them in wireless network designs.

Our research on layerless dynamic networks enables the characterization of fundamental performance regions without a presupposition of protocol stack layering. Previous results in layerless and cross-layer design (e.g. [9] and the references therein) typically involve designing higher-layer functions, including routing, scheduling, or resource allocation, according to the variation of the wireless physical channel via a joint optimization. In contrast, our layerless framework focuses on finding a simple interface to the physical layer that allows the upper layers to achieve optimal or near optimal cross layer performance based on the underlying channel conditions. We believe that this layerless design must exploit the unique broadcast features of wireless transmission through generalized network coding, including cooperation and relaying. Channel and network dynamics also require hierarchical codes that admit some outage in poor channel conditions or limit code blocklength under stringent delay constraints. Finally, our layerless design optimizes the use of noisy finite-rate feedback to achieve both capacity and robustness. We now describe in more detail our results and research directions in these areas.

Surely if the source and channel code designs cannot be optimally separated from each other in wireless channels, neither can be optimally separated from the network protocol stack governing end-to-end transmission that lies between them in wireless network designs.

Network Coding, Relaying, and Cooperation

The broadcast property is a very special feature of the wireless medium. When a node transmits, nodes other than the desired receiver can often overhear these transmissions. These listening nodes can use the information sensed from the channel to avoid collisions, sense and utilize unoccupied spectrum, remove interference via multiuser detection, dirty paper coding, or interference alignment, as well as *forward* overheard signals to their desired receiver(s) and/or other receivers to help them decode the message or remove interference [10]. The best of these possible strategies requires optimizing the performance tradeoffs of all network users and not just a single transmit-receive pair. Other design considerations include complexity of the cooperation, its scalability and stability, and the overhead cost to obtain the information required for cooperation.

Fundamental performance limits of the layerless approach have been investigated through the concept of network coding, which defines how source nodes embed information, intermediate nodes relay information, and destination nodes decode information. There has been significant recent activity on determining wireless network capacity regions based on network coding. Recent results indicate that *noisy network coding* can incorporate many cooperation and relaying strategies known to date [11]. Noisy network coding optimally treats interference: either

Our approach to incorporate delay and outage into our layerless network design is through the development of novel hierarchical channel codes where different encoded bits can be decoded under different time constraints and with different reliabilities.

(partially) decoding it or treating it as noise, depending on propagation conditions and the network topology. This coding yields the best known achievable rate regions for multisource multicast networks, matching or beating the achievable rates of deterministic network coding and erasure network coding. However, noisy network coding has not yet been applied to networks of more than a few nodes due to its complexity. At the other extreme, simple analog network coding, where a node forwards the signal it receives from all other users, has recently been shown to achieve capacity in the high power regime for certain network topologies. The simplicity of this technique has desirable scaling properties in large networks that do not hold for more complex network coding schemes. However, its performance is limited by propagated noise, and we expect this strategy would not perform well in noise-limited regimes.

Hierarchical Coding under Delay and Outage

One of the biggest limitations of Shannon capacity applied to wireless networks is the requirement that the probability of error must be asymptotically small, regardless of the channel conditions. For point-to-point channels with severe fading such as Rayleigh, this leads to a capacity of zero; in other words, Rayleigh fading is sufficiently severe so that no coding scheme can guarantee reliable communication for any finite SNR. A broader notion of capacity called *outage capacity* allows for some outage under poor channel and network conditions; outage capacity reflects common design practice where some loss of data is tolerated in exchange for higher overall data rates. In particular, allowing for some probability of outage can greatly increase communication rates during non-outage, since the encoding and decoding strategies need not be designed for worst-case conditions. Another limitation of Shannon capacity in the context of wireless networks is its inherent lack of delay constraints. As a result, capacity-achieving codes generally have asymptotically long blocklengths, which precludes incorporation of delay constraints into the analysis, although some recent results have made progress in developing tight performance bounds on finite-length codes [12].

Our approach to incorporate delay and outage into our layerless network design is through the development of novel *hierarchical* channel codes where different encoded bits can be decoded under different time constraints and with different reliabilities. Hierarchical codes typically require structure along with powerful short codewords due to stringent delay constraints. As part of our work we have characterized the fundamental performance limits of hierarchical codes in terms of the tradeoff between the rates, the reliability, and the delay of different submessages. Hierarchical codes are best used in the context of joint source-channel coding, which requires network-level awareness of bit requirements, leading to a joint source/channel/network code approach. With this approach, network control, protocol messages, as well as data with different quality-of-service requirements can all be treated in a uniform way while optimally utilizing limited network resources. The end-to-end performance benefits of these codes can be determined using the techniques described later by incorporating these codes into the physical layer models of the networks being optimized.

Optimal Use of Feedback

Feedback is ubiquitous in practical wireless network designs. One of the most common uses of feedback on point-to-point channels is to acknowledge when packets are received correctly (ACK) and to send retransmission requests (ARQ) when data is corrupted. Feedback is also used on these channels to send estimates of channel state information in order to adapt the link transmission policy to the link state. Channel access and routing protocols also exploit feedback information to establish connections and determine the least-congested routes. Yet the capacity of wireless channels and networks with feedback is largely unknown, raising the question as to whether these uses of feedback best exploit its potential to improve performance.

Intuitively, using feedback to convey ACKs and ARQs as well as channel and network state information (CSI and NSI, respectively) to transmitting nodes will both improve their performance and robustness as well as simplify their protocol design. However, the underlying time-varying nature of wireless channels and networks results in erroneous CSI and NSI being fed back. There have been thorough studies on the impact of the lack of CSI, or the lack of its precision, at either the transmitter or the receiver, on the capacity of wireless single-user, broadcast, and multiple access channels. Much less is known about the impact of imperfect or intermittent CSI on capacity of multihop networks and related questions such as how much overhead should be incurred to obtain and distribute CSI. This imperfect CSI not only affects optimality of signaling on each link, but also how routes are selected and the amount of interference caused to other nodes in the network. NSI can consist of node queue lengths, network connectivity, network dynamics, and other network characteristics. While performance of specific networks with imperfect NCI has been analyzed, the impact of this limited NSI on fundamental network performance limits is a wide-open research area. There is clearly much more work to be done in determining the value of CSI and NSI as well as the impact of their imperfect knowledge on wireless ad hoc network performance. But is feedback of ACKs, ARQs, CSI and NSI the optimal use of finite-rate noisy feedback in wireless networks? We have no capacity results to indicate that is the case. Thus, our work has also been exploring the fundamental performance limits of wireless networks where feedback can take any form subject to constraints on the feedback link. For example, we have developed results where feedback is

- Used by cognitive nodes to extract messages of other users to remove their interference and/or enhance their transmission.
- Exploited via structured code-trees to enhance capacity.

- Incorporated into transmission strategies to reduce delay and improve multiplexing-diversity tradeoff regions in multihop routes.

In addition to these results, a new perspective that applies well-developed stochastic control principles to determine capacity under feedback has emerged [13]. This perspective yields performance upper bounds for channels with feedback along with recursive feedback-based encoder designs. While the initial results are for point-to-point channels, the insights are applicable to general wireless networks. These stochastic control tools could become a powerful mechanism to address the impact of noise in the feedback channel and ultimately new insights on the optimal use of feedback in wireless networks.

Some Open Challenges

There is a vast amount of research that can be done on the topic of layerless dynamic networks, since any scheme can be imagined and then analyzed. The challenge is identifying the most promising approaches that yield significant performance improvements over existing methods and, ideally, approach the network performance upper bounds. The notion of generalized network coding, including ideal relaying and cooperation, holds much promise, and recent results indicate these techniques can be capacity-achieving for some specific network topologies and SNR regimes. However, these techniques do not scale well with network size, and simple scalable techniques like analog network coding are likely to perform well only in limited regimes. Thus, a significant open challenge is to develop optimal wireless network coding strategies that are scalable with reasonable complexity and perform much better than the standard techniques of amplify-forward, decode-forward, and compress-forward. Design and analysis of hierarchical codes must be extended from point-to-point links to multihop networks to determine their impact on delay and outage. Finally, creative and novel uses of feedback in networks, beyond feeding back noisy and compressed CSI/NSI as well as ACKs and ARQs, should be explored.

Application and Network Optimization

The third research area within our framework is to provide a universal algorithmic architecture that is capable of optimally balancing and trading off application requirements and network capabilities. The main idea is to find the operating point on the underlying network performance tradeoff region that optimizes a set of application or network performance metrics for all users in the network. The network performance tradeoff regions will typically coincide with wireless ad hoc network capacity regions, upper bounds or achievable regions, for example, those obtained using the techniques described in previous sections, but they could also consist of performance regions associated with existing systems and standards, or even be obtained from measured data in different environments. The optimal operating point is found algorithmically using tools from optimization and stochastic analysis. The approach places few restrictions on the network metrics and performance tradeoff regions it can optimize over. Moreover, when the chosen metric and performance regions are convex, global optimality can be assured.

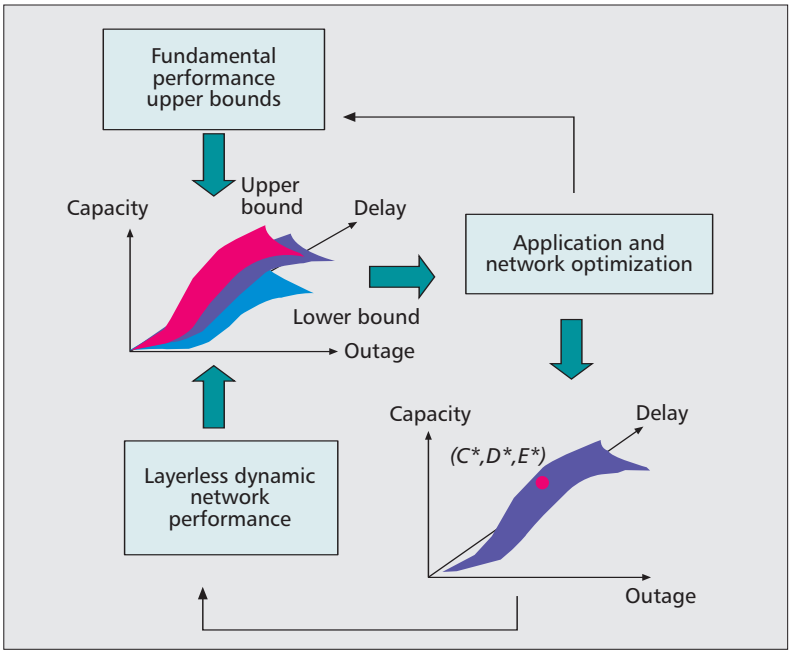

Figure 6. *Unification and synergies between research areas.*

This work complements and integrates with the other research areas within our overall framework since it indicates what a given network is capable of, how underlying network performance tradeoffs affect applications, and the performance gap associated with performance region upper and lower bounds in terms of application metrics. This unification and interplay between all three research areas within our proposed framework is illustrated in Fig. 6. Specifically, this figure shows a performance region indicating tradeoffs between capacity, delay, and outage. This is just one slice of the $n \times (n-1)$ performance region in an n-user network. The optimization algorithm will identify the optimal operating point on this tradeoff region to meet heterogeneous application requirements: some applications (e.g. file transfer) might require high data rates but have no stringent delay requirements, whereas others (e.g. voice on a handset) may have low data rate requirements but strict delay and energy constraints, whereas still others (video) might have both high rate and low delay requirements. Optimization determines how the underlying network resources can best be utilized to maximize the application-layer performance metrics associated with all these applications. Practical implementation of this architecture requires that the algorithms are distributed (i.e. each node acts on local information or obtains it with low communication overhead), adaptable to application and network dynamics, scalable in network size, asynchronous, and robust against interference and jamming. Our approach to this algorithmic architecture design

This generalization of MaxWeight is based in part on a workload relaxation obtained using cutset bounds. This workload relaxation makes a complex network simple, and therefore easy to optimize over, with tight bounds on the approximations associated with the relaxation.

combines new and existing theoretical and computational tools in optimization, control, stochastic network analysis, and game theory, as we now describe in more detail.

Dynamic Network Utility Maximization

Recent research in network optimization efficiently allocates resources among heterogeneous applications based on a metric of network utility (see [14] and the references therein). This approach, referred to as Network Utility Maximization (NUM), has led to a deeper understanding of wireline network architectures and the development of new protocols. However, it has several shortcomings when applied to wireless networks. In particular, the underlying premise of NUM is that network dynamics occur on time scales much longer than the algorithm convergence time, hence network structures are effectively static. Since rapidly changing conditions play a central role in wireless networks, a new *dynamic* framework is needed to optimize their performance.

We have developed a new dynamic NUM theory that combines NUM at the application layer with adaptation of resources and transmission policies at the physical and network layers. By optimizing performance at each protocol layer in terms of application metrics, significantly different transmission techniques can result. For example, physical-layer optimization no longer yields a water-filling power and rate allocation in flat fading. In addition, this framework adopts a multiperiod optimization to meet the performance requirements and hard-delay constraints of dynamic traffic. Dynamic NUM can perform utility maximization based on upper and lower performance region bounds obtained using the techniques described in previous sections. For example, we have used this framework to optimize resource allocation and network utility for the fading MAC capacity region relative to a general concave utility function of transmission rates with and without CSI.

While wireline NUM techniques and their variants lead to distributed implementations based on systematic decompositions of the optimization problem into locally solvable subproblems, this is often not the case for MANETs due to their shared physical channel constraints. In order to develop distributed dynamic NUM policies, we combine optimization methods with consensus policies. Consensus policies involve each node maintaining estimates of its own decision vector and updating it based on local information. Along these lines we have developed algorithms for distributed optimization of application performance metrics over time-varying MANETs [15]. These algorithms allow for optimization of general (convex) objective functions with time-varying local constraints, time-varying network connectivity, and time-varying underlying channel characteristics which may change faster than the convergence speed of the optimization algorithms.

Routing and Scheduling

Wireless network designs must also ensure that node queues remain small to minimize delay. One of the dominant approaches to queue-length-based control to ensure stable queues is the *MaxWeight* policy introduced in [16]. Under this policy, at each node information is stored at different queues depending on destination, and priority in transmission is given to traffic types that have the longest differential queue length. Although there has been much work in showing stability of such queue-length-based policies, this framework lacks any kind of mechanism for performance specification and optimization. We have introduced a new class of scheduling and routing policies within our research framework to deal with these shortcomings [17]. This new h-MaxWeight policy extends Maxweight from a policy based on a quadratic cost function to one based on more general h functions. Subject to a few technical conditions, the new policy is throughput-optimal for any network for which MaxWeight is throughput-optimal. Moreover, for appropriate choice of h the policy is approximately delay-optimal. This generalization of MaxWeight is based in part on a workload relaxation obtained using cutset bounds. This workload relaxation makes a complex network simple, and therefore easy to optimize over, with tight bounds on the approximations associated with the relaxation. In addition to h-MaxWeight, ideas from consensus policies can be used to obtain distributed algorithms that allow practical distributed implementation of generalized MaxWeight policies over wireless networks. Further, the use of the inherent geometry of the wireless network topology allows for simple, distributed iterative scheduling algorithms that have order-optimal delay.

Game Theory

One of the greatest appeals of wireless ad hoc networks is their distributed nature, which allows them to scale organically and be robust against single points of failure. In addition to the distributed optimization framework described above, a new distributed network control paradigm is emerging. In this approach "competitive situation-aware" users make autonomous decisions with regard to their network usage based on the current network conditions and their individual preferences [18]. These game-theoretic approaches usually aim to stabilize network operation through an equilibrium, where individual users cannot achieve better performance through individual actions. The basic premise of these approaches is that friendly competition among users may lead to globally-optimal distributed control policies based on local information that are practical to implement.

Game theory naturally addresses competition among users for limited wireless network resources. In our approach to this problem, we exploit a "law of large numbers" for dynamic game theoretic models of interaction between multiple devices [19]. In the limit where the number of devices becomes large, we obtain a computationally tractable model that allows us to study the efficiency properties of proposed spectrum management algorithms in dynamic environments. The game theoretic approach is also useful in sub-network selection when multiple users share a common network characterized

by a limited resource such as power. For example, we can show that a game theoretic approach enables multiple unicasts to transmit messages through a shared network in a manner that minimizes the total number of transmissions (and therefore power) required to meet the combined goals of all users. Network topology formation is another problem where a game-theoretic formulation captures the tradeoffs between network metrics such as connectivity and the overhead cost of routing and link maintenance for network nodes. Another protocol where it is natural to apply a game-theoretic approach is competitive scheduling. When posed in a game-theoretic context, competitive scheduling can incorporate different user objectives and channel state processes, and sheds light on the equilibrium characterization and distributed convergent dynamics for competitive versus coordinated scheduling among multiple users.

Some Open Challenges

The tools of optimization, control and game theory are ideal for tackling the complex design and performance analysis challenges of large networks. Many of the results in these areas assume centralized control, and developing distributed algorithms remains a challenge for much of this work. Moreover, research in network and application optimization has mostly failed to incorporate the strategies and insights obtained via information-theoretic analysis on network performance bounds. Steps in this direction are just beginning — this integration is challenging as the information-theoretic tools are only available for small networks, and often these techniques make the network optimization problems nonconvex. Network equivalence and similar techniques that abstract away the details of the physical layer while maintaining capacity-achieving strategies at that layer may facilitate progress in this direction. Consummating the union between all three research areas within our framework holds the promise for significant breakthroughs in determining fundamental performance bounds on wireless networks.

Summary

While much progress has been made on determining the fundamental performance limits of wireless channels, wireless network design above the physical layer still lacks a fundamental theory to guide it. We have presented a research framework to make progress in both determining the fundamental performance limits of wireless ad hoc networks, as well as developing design insights and networking techniques to approach these limits. Our interdisciplinary approach integrates tools and methodologies from traditional Shannon Theory along with network theory, combinatorics, optimization, control, and game theory. Preliminary results have already led to significant performance gains as well as new wireless networking techniques, protocols, and insights. Further progress will entail integration of the results, ideas, and tools across all three research areas within our framework. The goal of this article was to describe the many open problems in this area, set forth our approach to this great challenge, and inspire other researchers to join our quest toward a theory of fundamental performance limits for wireless networks that is broad, interdisciplinary, and leads to significant breakthroughs in both theory and practice.

The tools of optimization, control and game theory are ideal for tackling the complex design and performance analysis challenges of large networks. Many of the results in these areas assume centralized control, and developing distributed algorithms remains a challenge for much of this work.

Acknowledgments

The authors gratefully acknowledge the support of the DARPA ITMANET program, as well as the valuable insights, suggestions, and encouragement of the current and former program managers Aaron Lazarus and J. Christopher Ramming. The participation of Ananthram Swami and Richard Barron as advisors to the program has also been invaluable, as well as the lively discussions and debates with the other ITMANET project team focused on Non-equilibrium Information Theory. Finally, we gratefully acknowledge the entire team of Principle Investigators, students, postdocs and advisors whose research and insights have contributed to the development of the research framework described in this article. The evolution of our approach to this challenging problem would not have been possible without them.

References

[1] A. Goldsmith *et al.*, "Capacity Limits of MIMO Channels," *IEEE JSAC*, vol. 21, no. 5, June 2003, pp. 684–702.
[2] F. Xue and P. R. Kumar, "Scaling Laws for Ad-Hoc Wireless Networks: An Information Theoretic Approach," *NOW J. Foundations and Trends in Networking*, vol. 1, no. 2, 2006, pp. 145–270.
[3] V. Cadambe and S. A. Jafar, "Interference Alignment and Degrees of Freedom of the k-User Interference Channel," *IEEE Trans. Info. Theory*, vol. 54, no. 8, 2009, pp. 3425–41.
[4] B. Hajek and T. Ephremides, "Information Theory and Communications Networks: An Unconsummated Union," *IEEE Trans. Info. Theory*, Oct. 1998, pp. 2416–34.
[5] J. Andrews *et al.*, "Rethinking Information Theory for Mobile Ad Hoc Networks," *IEEE Commun. Mag.*, 2008, pp. 94–101.
[6] P. Moulin, "Towards Strong Converses in MANETs," *Proc. IEEE Int'l. Symp. Info. Theory*, June 2009, pp. 1958–62.
[7] R. Koetter, M. Effros, and M. Médard, "A Theory of Network Equivalence, Parts I and II". Part I appeared in the *IEEE Trans. Inf. Theory*, Feb. 2011, pp. 972-995. Part II is under review in the same journal and can be found at http://arxiv.org/abs/1007.1033v2. Portions of these works appeared in the *Proc. Allerton Conf. Communication, Control, and Computing*, Sept. 2009 and in the *IEEE Info. Theory Wksp.*, June 2009.
[8] A. S. Avestimehr, S. Diggavi, and D. N. C. Tse, "A Deterministic Approach to Wireless Relay Networks," *IEEE Trans. Inf. Theory*, Apr. 2011, pp. 1872-1905.
[9] V. Srivastava and M. Motani, "Cross-Layer Design: A Survey and the Road Ahead," *IEEE Commun. Mag.*, vol. 12, Dec. 2005, pp. 112–19.
[10] A. Goldsmith *et al.*, "Breaking Spectrum Gridlock with Cognitive Radios: An Information Theoretic Perspective," *IEEE Proc.*, vol. 97, May 2009, pp. 894–914.
[11] S. H. Lim *et al.*, "Multi-Source Noisy Network Coding," *Proc. Int'l. Symp. Info. Theory*, vol. 95, no. 1, pp. 604–08, 2010, also submitted to the IEEE Trans. Info. Theory and to Arxiv, http://arxiv.org/abs/1002.3188.
[12] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Info. Theory*, vol. 56, no. 5, 2009, pp. 2307–59.
[13] T. P. Coleman, "A Stochastic Control Viewpoint on 'Posterior Matching-Style' Communication Schemes," *Proc. IEEE Intl. Symp. Info. Theory*, July 2009, pp. 1520–24.
[14] M. Chiang *et al.*, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proc. IEEE*, vol. 95, no. 1, 2007, pp. 255–312.

Preliminary results have already led to significant performance gains as well as new wireless networking techniques, protocols, and insights. Further progress will entail integration of the results, ideas, and tools across all three research areas within our framework.

[15] I. Lobel and A. Ozdaglar, "Distributed Subgradient Methods in Random Networks," *Proc. Allerton Conf. Commun. Control., Comp.*, Sept. 2008, also to appear in IEEE Trans. Automatic Control.

[16] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queuing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," *IEEE Trans. Automatic Control*, vol. 37, no. 12, 1992, pp. 1936–48.

[17] W. Chen *et al.*, "Coding and Control for Communication Networks," *Queuing Systems: Theory and Applications*, vol. 63, no. 1–4, Dec. 2009, pp. 195–216.

[18] R. Johari and R. Berry, *Game Theory in Networks*, NoW Publishers, 2009.

[19] S. Adlakha *et al.*, "Oblivious Equilibrium for Large-Scale Stochastic Games with Unbounded Costs," *Proc. IEEE Conf. Dec. Contrl.*, Dec. 2008, pp. 1092–98.

BIOGRAPHIES

ANDREA GOLDSMITH [F] (andrea@wsl.stanford.edu) is a professor of Electrical Engineering at Stanford University. She also founded Quantenna Communications and served as its CTO, and has held several other industry positions. She is a Fellow of Stanford. She has received several awards for her research, including the joint paper award from the IEEE Communications and Information Theory Societies, the IEEE Wireless Communications Committee Technical Achievement Award, and the Silicon Valley/San Jose Business Journal's Women of Influence Award. She has co-authored two books on wireless communications, and has served as editor, on the Board of Governors, and as a Distinguished Lecturer for both the IEEE Communications and Information Theory Societies.

MICHELLE EFFROS [F] is a Professor of Electrical Engineering at the California Institute of Technology. She is a recipient of the NSF CAREER Award, Charles Lee Powell Foundation Award, and Richard Feynman-Hughes Fellowship. In 2002, she was cited by Technology Review as one of the world's top 100 young innovators. She was co-recipient of the 2009 Joint Paper Award of the Communications and Information Theory Societies. She currently serves on the Board of Governors for the IEEE Information Theory Society and on the Advisory Committee for the National Science Foundation Directorate for Computer and Information Science and Engineering.

RALF KOETTER was on the faculty of the University of Illinois at Urbana-Champaign from 1999 to 2007 and served as Head of Institute at the Institute for Communications Engineering, Technical University of Munich from 2007 to 2009. He received an IBM Invention Achievement Award in 1997, an NSF CAREER Award in 2000, an IBM Partnership Award in 2001, and the Vodafone Innovationspre is in 2008. He was co-recipient of the 2004 and 2010 Paper Awards of the IEEE Information Theory Society and the 2009 Joint Paper Award of the IEEE Communications and Information Theory Societies.

MURIEL MÉDARD [F] is a Professor in EECS at MIT. She received five degrees from MIT. She has been associate or guest editor for numerous journals and TPC chair or member for numerous conferences. She received the 2009 Communication Society and Information Theory Society Joint Paper, the 2009 William R. Bennett Prize in the Field of Communications Networking, and the 2002 IEEE Leon K. Kirchmayer Prize Paper Award. She received a NSF Career Award in 2001, the 2004 MIT Harold E. Edgerton Faculty Achievement Award, and was named a Gilbreth Lecturer by the National Academy of Engineering in 2007.

ASU OZDAGLAR [M'95] received her Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 2003. Since 2003, she has been a faculty member in the Electrical Engineering and Computer Science Department at the Massachusetts Institute of Technology, where she is currently the Class of 1943 Associate Professor. Her research interests include optimization theory, game theory, with applications in communication and social networks, and distributed optimization and control. She is the recipient of a Microsoft fellowship, the MIT Graduate Student Council Teaching award, the NSF Career award, and the 2008 Donald P. Eckman award.

LIZHONG ZHENG received the B.S and M.S. degrees, in 1994 and 1997 respectively, from the Department of Electronic Engineering, Tsinghua University, China, and the Ph.D. degree, in 2002, from the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Since 2002, he has been working in the Department of Electrical Engineering and Computer Sciences, where he is currently an associate professor. His research interests include information theory, wireless communications and wireless networks. He received IEEE Information Theory Society Paper Award in 2003, an NSF CAREER award in 2004, and the AFOSR Young Investigator Award in 2007. He is currently an associate editor of IEEE Transactions on Information Theory.