Network Coding and Reliable Communications

Academic and Research Staff
Professor Muriel Medard
Dr. Chris Ng (postdoctoral fellow)
Dr. Atilla Eryilmaz (postdoctoral fellow)
Dr. Una-May O'Reilly (research scientist)

Visiting Scientists and Research Affiliates
Mr. Toshihito Fujiki (Sony Corporation)
Ms. Nadia Fawaz (Eurecom./ Supelec)
Professor Joao Barros (University of Porto)
Mr. Gerhard Maierbacher (University of Porto)
Ms. Luisa Lima (University of Porto)
Mr. Danail Traskov (Technical University of Munich)
Mr. Paulo Vasco Falcao Martins De Oliveira (University of Porto)
Mr. Babak Anthony Nazer (Unoversity of California at Berkeley)
Mr. Jaspreet Singh (University of California Santa Barbara)
Mr. Javier de Pedro Carracedo
Mr. Joao-Paulo Vilela (University of Porto)
Dr. Ivana Maric (Stanford)

Graduate Students
Mr. Vishal Doshi
Mr. Sheng Jing
Ms. Minji Kim
Mr. Minkyu Kim
Mr. Daniel Lucani
Mr. Ali ParandehGheibi
Ms. Shirley Shi
Mr. Guy Weichenberg
Ms. Fang Zhao

Professor Médard's group works extensively on capacity, in collaboration with CSAIL, LIDS, Caltech, the University of Illinois Urbana-Champaign (UIUC), UCLA, Stanford and the Technical University of Munich (TUM). Network coding provides cost benefits in a variety of settings, for instance, wireless networks, where the cost may be measured in expended energy, or wireline networks, where they reduce congestion. In the area of network coding for wireless networks, she, Professor Katabi and Professor Ozdaglar have a DARPA contract through BAE to apply network coding to mobile ad-hoc networks (MANET) – CBMANET CONCERTO. The first phase of this contract has been very successful and this next phase concentrates on technology transfer to the warfighter. She has considered algorithmic issues of network coding in MANETs through the Army Research Office DAWN program, for which she is the sole MIT PI, in collaboration with University of California Santa Cruz, Stanford, University of Maryland College Park and University of California Los Angeles. In the area of network coding security, she is PI of a DARPA program and of an AFOSR program with Caltech and the University of Illinois Urbana-Champaign for the application of network coding to protect data under eavesdropping and Byzantine attacks. She is also MIT PI of a DARPA program through BAE for intrinsically assured MANETs (IAMANET PIANO), with University of Massachusetts Amherst, Stanford and University of Texas at Austin. She investigates the general applicability of network coding to wireless systems through a NSF contract with Professor Katabi on XORs in the air.

In the area of information theory, Professor Medard has obtained, with Professors Ozdaglar, Shah and Zheng, a DARPA contract for the study of information theory for MANETs (ITMANET) with UIUC, TUM, Stanford and Caltech. This work has led to work on multiple access power

control, distributed functional compression and analog network coding. She has also investigated fundamental coding issues in network coding through an NSF ITR project with Caltech, UIUC and Alcatel/Lucent Bell Laboratories.

Professor Medard also works in the area of optical network performance, reliability and robustness. With Professor Chan, she conducts research in the area of optical network capacity and optical access networks through an NSF FIND contract. Under the DARPA FONA project for optical networks, she and Professor Chan are investigating the limits of reliability of optical networks.

Professor Medard has also explored new areas at the intersection of communications and biochemistry for genomics and spectroscopy in collaboration with the Broad Institute and the Department of Chemistry.

1. Information Theory for MANETs

Sponsor:
DARPA ITMANET under the FLOWS project

The purpose of this project is to investigate the information-theoretic limits of MANETs. The intrinsic limits consider topology, bandwidth, delay, capacity and energy. Figure 1 below illustrates the main metrics of this project.

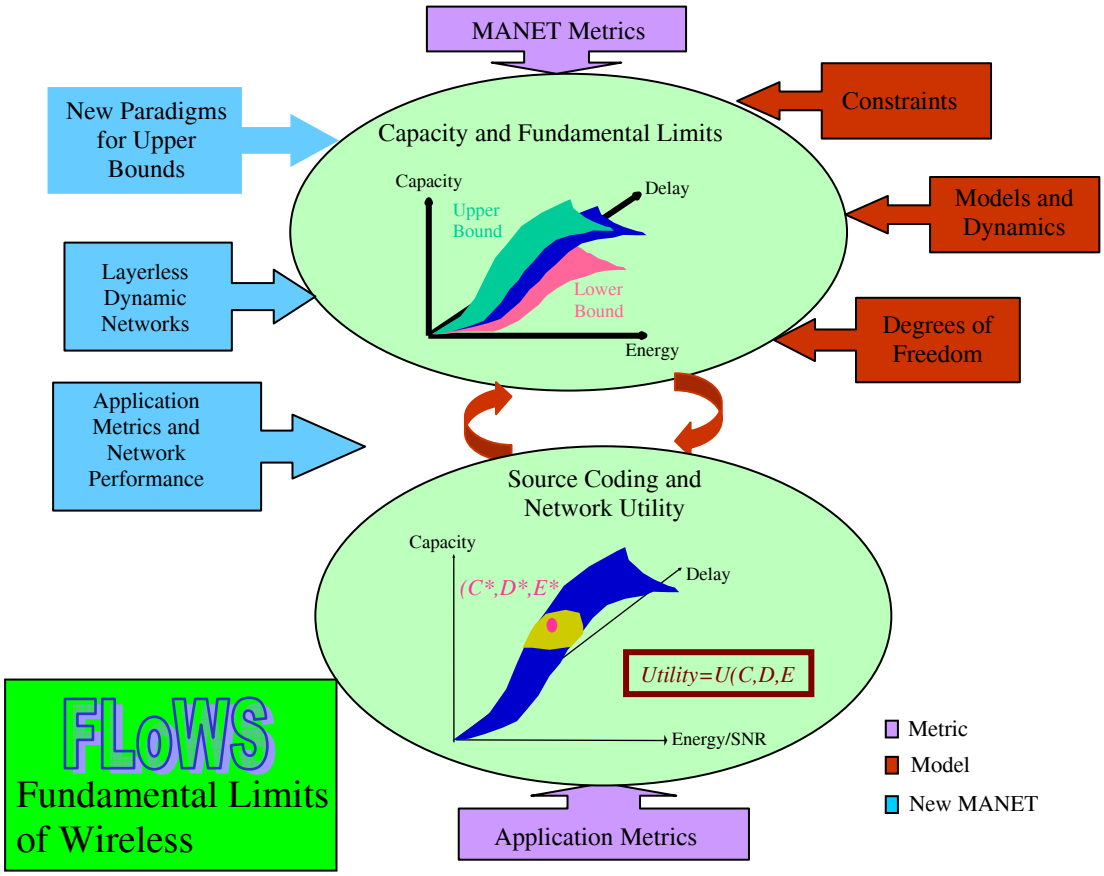


Figure 1: main thrusts of th FLOWS project and its relation to metrics.

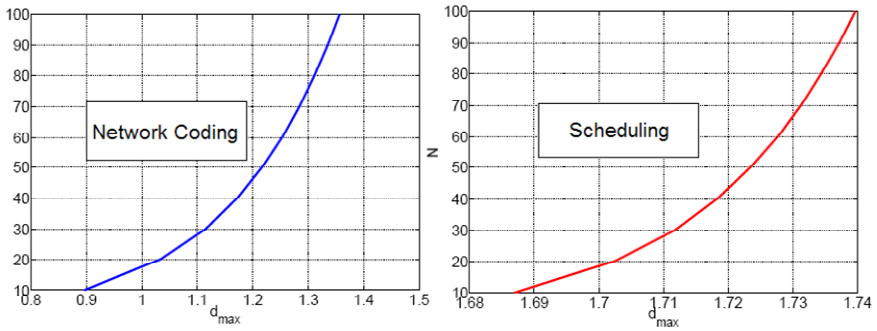
One of the effects of channel dynamics is that feedback can be used to adapt to changing conditions, Transmitter knowledge of channel state has a great impact on wideband fading channel capacity. However, in the low SNR regime, power per dimension does not suffice to provide an accurate measurement of the channel over the entire spectrum.

In the presence of feedback, we may collect information at the transmitter about some aspects of the channel quality over a certain portion of the spectrum. In this work, we investigate the effect of such information. With Professor Zheng and our student consider channel testing with a finite amount of energy over a block-fading channel in both time and frequency. We consider a transmission scheme in which the wideband channel is decomposed into many parallel narrowband subchannels, each used with a binary modulation scheme. The quality of each subchannel corresponds to the crossover probability of a binary symmetric channel. We use a multi-armed bandit approach to consider the relative costs and benefits of allotting energy for testing versus transmission, and for repeated testing a single subchannel versus testing different subchannels. We give both upper and lower bounds on the number of subchannels that should be probed for throughput maximization under the scheme we have chosen. Our bounds are in terms of available transmission energy, available bandwidth and fading characteristics of the channel. Moreover, in our numerical results, the two bounds are close.

Power efficiency is a capital issue in the study of mobile wireless nodes owing to constraints on their battery size and weight. In practice, especially for low-power nodes, it is often the case that the power consumed for non-transmission processes is not always negligible. With Professor Zheng and our student, we have considered the channels with a special form of overhead: a processing energy cost whenever a non-zero signal is transmitted. We have shown that under certain conditions, achieving the capacity of such channels requires intermittent, or `bursty', transmissions. Thus, an optimal sleeping schedule can be specified for wireless nodes to achieve the optimal power efficiency. We have shown that in the low SNR regime, there is a simple relation between the optimal burstiness and the overhead cost: one should use a fraction of the available degrees of freedom at an SNR level of $(2\varepsilon)^{1/2}$, where ε is the normalized overhead energy cost. We have extended this result to use bursty Gaussian transmissions in multiple parallel channels with different noise levels. Our result can be intuitively interpreted as a "glue pouring" process, generalizing the well-known water pouring solution. We have used this approach to compute the achievable rate region of the multiple access channel with overhead cost.

The issue of delay in an information-theoretic setting has always been rendered difficult by the fact that traditional notions of capacity have no upper bound on delay. In order to quantify the delay benefits of coding in MANETs, we have investigated rateless codes with Professor Ozdaglar, Professor Eryilmaz (Ohio Sate University) and our student. In an unreliable packet network setting, we study the performance gains of optimal transmission strategies in the presence and absence of coding capability at the transmitter, where performance is measured in delay and throughput. Although our results apply to a large class of coding strategies including Maximum Distance Separable (MDS) and Digital Fountain codes, we use random network codes in our discussions because these codes have a greater applicability for complex network topologies. To that end, after introducing a key setting in which performance analysis and comparison can be carried out, we provide closed form as well as asymptotic expressions for the delay performance with and without network coding. We show that the network coding capability can lead to arbitrarily better delay performance as opposed to traditional strategies. Figure 2 below illustrates this benefit in the case of inelastic traffic.

- **Delay scaling for inelastic traffic**
- Inelastic Traffic: each user has a delay constraint associated with it and is admitted only if the mean waiting time is lower than its constraint
- Scaling laws cannot be applied here
- Each user has a delay constraint Θ that is distributed uniformly between θ and d_{max}



Number of supportable users as a function of delay constraints for random network coding and round robin for K=20

Figure 2: delay gais obtained from network coding

With Professor Ozdaglar, Professor Eryilmaz (Ohio State University) and our student, we have considered the problem of rate allocation in a fading Gaussian multiple-access channel with fixed transmission powers. The goal is to maximize a general concave utility function of the expected achieved rates of the users. There are different approaches to this problem in the literature. From an information theoretic point of view, rates are allocated only by using the channel state information. The queueing theory approach utilizes the global queue-length information for rate allocation to guarantee throughput optimality as well as maximizing a utility function of the rates. In this work, we have made a connection between these two approaches by showing that the information theoretic capacity region of a multiple-access channel and its stability region are equivalent. Moreover, our numerical results show that a simple greedy policy which does not use the queue-length information can outperform queue-length based policies in terms of convergence rate and fairness.

2. Information-theoretic aspects of network coding

Sponsor:
NSF ITR network coding project.

We have considered, with Professor Shah and Professor Koetter from TUM and our student, the problem of serving multicast flows in a crossbar switch is considered. We have shown linear network coding across packets of a flow can achieve a larger rate region compared to the no-coding case. In addition to such throughput gains, the characterization of the rate region becomes simpler when coding is allowed. We have characterized he rate region with coding graph-theoretically, in terms of the stable set polytope of the “enhanced conflict graph” of the traffic pattern. No such graph theoretic characterization of the rate region is known for the case of fanout splitting without coding. The minimum speedup needed to achieve 100% throughput with coding we have shown to be upper bounded by the imperfection ratio of the enhanced conflict graph. In particular, we have applied this result to K x N switches with traffic patterns consisting of unicasts and broadcasts only and an upper bound is obtained on the speedup. Such bounds

show that in multicast switches, speedup, which is usually implemented in hardware, can often be substituted by network coding, which can be done in software. We have shown that computing an offline schedule (using prior knowledge of the flow arrival rates) can be reduced to certain graph coloring problems. We have also proposed a graph-theoretic online scheduling algorithm (using only current queue occupancy information), that stabilizes the queues for all rates within

3. Evolutionary methods in network coding applied to MANETs

Sponsor:
ARO DAWN program

While network coding has been shown to offer various advantages over traditional routing, in order to see network coding widely deployed in real networks, it still remains to show that the amount of overhead incurred by additional coding operations can be kept minimal and eventually outweighed by the benefits network coding provides.

Whereas most network coding solutions assume that the coding operations are performed at all nodes, we have pointed out in our previous work that it is often possible to achieve the network coding advantage by coding only at a subset of nodes. Determining a minimal set of nodes where coding is required is NP-hard, as well as finding its close approximation. Thus, we have previously proposed evolutionary approaches toward a practical multicast protocol that achieves the full benefit of network coding in terms of throughput, while performing coding operations only when required at as few nodes as possible. Suppose that we have an ad hoc wireless network currently operating solely with traditional routing and wish to employ network coding on the network to achieve the maximal throughput promised by the network coding theory. However, for handling of the mathematical operations required for coding, the network nodes that have to perform network coding may need some changes in their software and/or hardware, which would necessarily incur some additional cost. Therefore, a very interesting and also practically important question that may arise in such a situation is whether we should change the entire nodes in the network for network coding or modifying only a subset of nodes would be enough. Along the same direction, many more interesting questions may follow: If only a small number of coding nodes are enough, as previously found in [1, 3] with a generic network model, where should those coding nodes be located? Can we fix their locations despite varying communication demands? How should those coding nodes interact with other non-coding nodes?

In our recent work, we considered the multicast scenario in a heterogeneous ad hoc wireless network where a number of coding nodes are to be placed among the legacy nodes that do not handle network coding operations well, to which our previously proposed evolutionary framework is applied to provide better understanding, if not complete answers, of the questions raised above. In particular, we have shown in [6] that our evolutionary approach is well generalized to the case of heterogeneous wireless networks with only slight changes in the fitness function and the backward evaluation phase, retaining its key advantages over existing centralized approaches [7, 8] in terms of the efficient operation through its spatially and temporally distributed structure and the superior performance in finding a minimal set of coding nodes.

Once a set of coding nodes is given, we can run our distributed algorithm directly over the network, during which each legacy node temporarily emulates coding operations on the application layer to find a minimal set of the legacy nodes that have to keep performing the emulated coding during the normal operation of the network to achieve the multicast capacity. More interestingly, we can utilize our evolutionary approach to investigate various issues regarding where and how to place coding nodes in heterogeneous wireless networks, which led to the following findings:

• To find out how many coding nodes are required to achieve the multicast capacity, we generated 100 random wireless topologies for each of 18 different parameter sets with either 20

or 40 nodes and the number of receivers varying from 4 to 10. We found that in most cases (over 90% for networks with 20 nodes and from 57% to 86% for networks with 40 nodes) network coding is not needed at all to achieve the multicast capacity. Even in the cases where network

coding is needed, the number of coding nodes does not exceed 15% of the total number of nodes; the number of required coding nodes is at most 3 in networks with 20 nodes and 5 in those with 40 nodes. This suggests that, while it is necessary to assume network coding everywhere initially to calculate the multicast capacity, to actually achieve the calculated maximal throughput coding operations may not be needed at all, and even when needed, only at a very small subset of nodes, thus incurring very little amount of overhead. Figure 3 below illustrates some of these results.

Nodes	Sinks	Rate	No. of Coding Nodes					
			0	1	2	3	4	5
20	4	4	96	3	-	1	-	-
		6	97	3	-	-	-	-
	6	4	93	7	-	-	-	-
		6	95	5	-	-	-	-
	8	4	96	4	-	-	-	-
		6	91	7	2	-	-	-
40	4	4	82	17	1	-	-	-
		6	79	14	5	1	1	-
		8	86	11	2	1	-	-
	6	4	74	19	6	1	-	-
		6	66	22	7	4	1	-
		8	83	12	4	1	-	-
	8	4	68	26	5	1	-	-
		6	62	27	7	3	1	-
		8	65	25	7	2	-	1
	10	4	72	21	6	1	-	-
		6	57	30	5	5	2	1
		8	68	20	9	2	1	-

Figure 3: number of topologies, out of 100 random ones, for which the calculated minimum number of coding nodes is as specified.

• With a slight change in the fitness function, we have found that, for the coding nodes found above, it is often possible to find alternative nodes to perform coding operations to achieve the given multicast capacity. In other words, the location of the coding nodes can often be flexible rather than fixed for a specific traffic pattern, implying that the found coding nodes can be shared by different multicast requests.

• To verify the intuition of sharing coding nodes among different communication demands, we picked two representative topologies from those used in the first experiment, and for each topology we randomly generated another 30 multicast requests to find the locations of the coding

nodes required for different traffic patterns. We found out that at least half of the coding nodes were in common with different communication demands for the both topologies. This indicates that placing coding nodes at the commonly found locations for a number of sampled traffic patterns may be a good strategy in practice.

• It is worth to point out that at the end of the iteration, our algorithm yields not only the set of the coding nodes, but also the relevant network code at each interior node, whether it indicates coding or routing. Therefore, the problem of interactions among coding and non- coding nodes is already dealt with implicitly within the framework of our algorithm, whether the algorithm is used to minimize the number of legacy nodes that have to emulate coding with the given set of coding nodes or to find the potential locations of the coding nodes. In addition, given the vulnerability of ad hoc wireless networks to various kinds of losses, we have shown that, thanks to its iterative nature, our algorithm can operate without much disruption even in the presence of a moderate level of packet erasures caused by various reasons. Furthermore, our algorithm may become even more robust by employing the temporally distributed structure in our previous work, whose initial motivation was better utilization of computational resources over the network. We have shown that the temporally distributed structure also offers a significant advantage in overcoming the adverse effect of packet erasures on the performance of the algorithm.

4. Practical approaches to wireless network coding

Sponsors:
NSF XORs in the air
DARPA CBMANET

Wireless networks suffer from interference and, in some cases, considerable delay. We have considered how to create practical schemes that allow us to design network coding mechanisms in the context of wireless settings, so that physical layer issues are explicitly taken into account in the development of our codes. Such issues are of particular importance in MANETs, where the paucity of resources, the variability of the topology and the uncertainty in the channels render the physical layer effects particularly challenging.

Network coding has been shown to improve throughput and reliability in a variety of theoretical and practical settings. But it has had limited success in areas like sensor networks due to its two limitations. First, network codes are ”all-or-nothing” codes; the sink cannot decode any information unless it receives as many coded packets as the original number of packets. Second, sensor networks often measure physical signals which show a high degree of spatial correlation; present network coding techniques cannot perform in-network lossy compression to take advantage of the spatial correlation. With Professor Katabi, Professor Jaggu (Chinese University of Hong Kong) and students, we have presented ”Real” Network Codes that are linear over real fields. We build on recent results from Compressed Sensing to develop new codes which can be decoded to get progressively more accurate approximations as more coded packets are received at the sink. Further, they can compress distributed correlated data inside the network without requiring that the nodes know how the data is correlated. Thus, Real Network Codes combine two exciting but hitherto separate areas, Network Coding and Compressed Sensing, allowing them to keep the advantages of network coding, but also make them capable of finding low distortion approximations with partial information and perform distributed compression of correlated data.

We have furthered our consideration of the interplay of the physical layer and network coding by considering symbol-level network coding. With Professor Katabi, her student and Professor Balakrishnan, we have introduced FUSE, a system that improves the throughput of wireless mesh networks. FUSE exploits a basic property of mesh networks: even when no node receives a packet correctly, any given bit is likely to be received by some node correctly. Instead of insisting on receiving correct packets, FUSE routers use physical layer hints to make their best

guess about which bits in a corrupted packet are likely to be correct and forward them to the destination. Even though this approach inevitably lets erroneous bits through, we find that it can achieve high throughput without compromising end-to-end reliability. The core component of FUSE is a novel network code that operates on small groups of bits, called symbols. It allows the nodes to opportunistically route groups of bits to their destination with low overhead. FUSE's network code also incorporates an end-to-end error correction component that the destination uses to correct any errors that might seep through. We have implemented FUSE on a software radio platform running the Zigbee radio protocol. Our experiments on a 25-node indoor testbed show that FUSE has a throughput gain of a factor of 2.8 over MORE, a state-of-the-art opportunistic routing scheme, and about 3.9 times over traditional routing using the ETX metric. Figure 4 below illustrates some of the throughput advantages of FUSE.

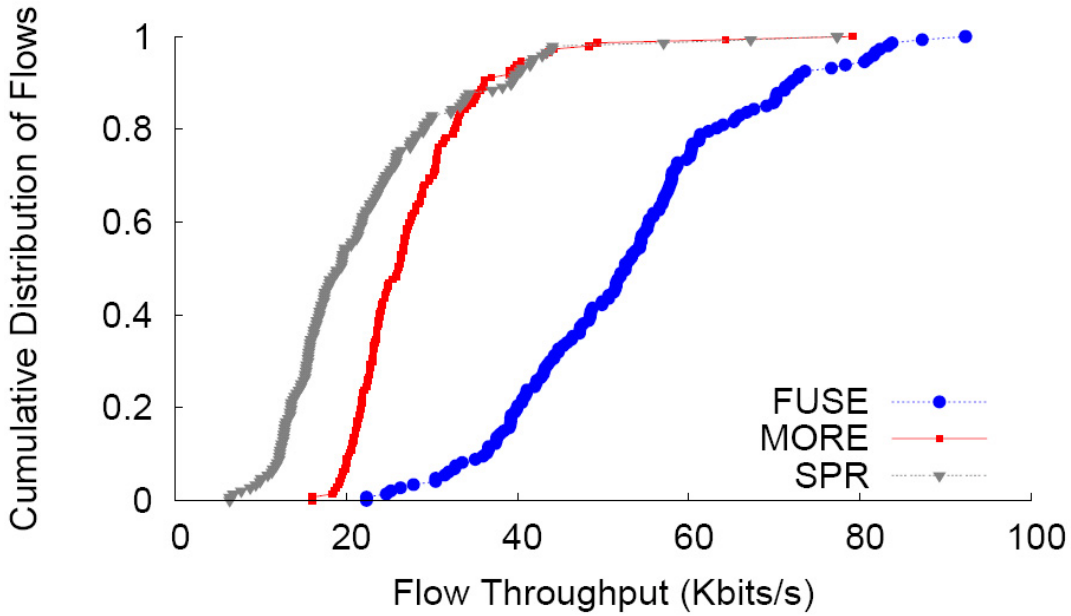


Figure 4: The figure shows that FUSE has a median throughput gain of 2.1 times over MORE, the state-of-the-art packet level opportunistic routing protocol, and 2.9× over SPR, a single path routing protocol based on the ETX metric.

The overhearing of transmissions that is inherent in wireless communications has generally been considered as a deleterious effect, leading to interference. With Professor Katabi, Professor Crowcroft (University of Cambridge) and their students, we have proposed COPE, a new architecture for wireless mesh networks. In addition to forwarding packets, routers mix (i.e., code) packets from different sources to increase the information content of each transmission. We show that intelligently mixing packets increases network throughput. Our design is rooted in the theory of network coding. Prior work on network coding is mainly theoretical and focuses on multicast traffic. This paper aims to bridge theory with practice; it addresses the common case of unicast traffic, dynamic and potentially bursty flows, and practical issues facing the integration of network coding in the current network stack. We evaluate our design on a 20-node wireless network, and discuss the results of the first testbed deployment of wireless network coding. The results show that using COPE at the forwarding layer, without modifying routing and higher layers, increases network throughput. The gains vary from a few percent to several folds depending on the traffic pattern, congestion level, and transport protocol.

The variability of MANETs entails considerable difficulties in reacting to changing conditions and also in acquiring global information about the network. The problem of establishing minimum-cost multicast connections in coded networks can be viewed as an optimization problem, and decentralized algorithms were proposed by Lun et al. to compute the optimal subgraph using the subgradient method on the dual problem. However, the convergence rate problem for these

algorithms remains open. There are limited results in the literature on the convergence rate of the subgradient method in the dual space, but the convergence rate of the primal solutions was not known. With Professor Ozdaglar and our student, we have analyzed the convergence rates of the min-cost subgraph algorithms in both the dual and the primal spaces. We have shown that using the incremental subgradient method on the dual problem with appropriately chosen step sizes yields linear convergence rate to a neighborhood of the optimal solution. Also, if we use constant step sizes in the subgradient method and simple averaging for primal recovery, the primal solutions recovered from the dual iterations converge to a neighborhood of the optimal solution with rate O(1/n).

Another important aspect of wireless communications is latency, particularly in acoustic channels or channels over long distances such as satellite channels. With Dr. Stojanovic and our student, we have considered underwater acoustic channels. Our goal of this research is two-fold. First, to establish a tractable model for the underwater acoustic channeluseful for network optimization in terms of convexity. Second, to propose a network coding based lower bound for transmission power in underwater acoustic networks, and compare this bound to the performance of several network layer schemes. The underwater acoustic channel is characterized by a path loss that depends strongly on transmission distance and signal frequency. The exact relationship among power, transmission band, distance and capacity for the Gaussian noise scenario is a complicated one. We have provided a closed-form approximate model for 1) transmission power and 2) optimal frequency band to use, as functions of distance and capacity. The model is obtained through numerical evaluation of analytical results that take into account physical models of acoustic propagation loss and ambient noise. We have applied network coding to determine a lower bound to transmission power for a multicast scenario, for a variety of multicast data rates and transmission distances of interest for practical systems, exploiting physical properties of the underwater acoustic channel. The results quantify the performance gap in transmission power between a variety of routing and network coding schemes and the network coding based lower bound. We illustrate results numerically for different network scenarios.

5. Security aspects of network coding.

Sponsors:
DARPA IAMANET program
AFOSR network coding security program

Network coding substantially increases network throughput. But since it involves mixing of information inside the network, a single corrupted packet generated by a malicious node can end up contaminating all the information reaching a destination, preventing decoding. With Professors Jaggi (Chinese University of Hong Kong), Katabi, Effros (Caltech), Langberg (Open University of Israel) and our students, we have introduced distributed polynomial-time rate-optimal network codes that work in the presence of Byzantine nodes. We have developed algorithms that target adversaries with different attacking capabilities. When the adversary can eavesdrop on all links and jam z links, our first algorithm achieves a rate of C − 2z, where C is the network capacity. In contrast, when the adversary has limited eavesdropping capabilities, we provide algorithms that achieve the higher rate of C − z. Our algorithms attain the optimal rate given the strength of the adversary. They are information theoretically secure. They operate in a distributed manner, assume no knowledge of the topology, and can be designed and implemented in polynomial-time. Furthermore, only the source and destination need to be modified; non-malicious nodes inside the network are oblivious to the presence of adversaries and implement a classical distributed network code. Finally, our algorithms work over wired and wireless network.

Network coding can also be used to enhance cryptographic approaches. With Professor Barros (University of Porto) and his students, we have considered the issue of confidentiality in multicast network coding, by assuming that the encoding matrices, based upon variants of random linear

network coding, are given only to the source and sinks. Based on this assumption, we provide a characterization of the mutual information between the encoded data and the two elements that

can lead to information disclosure: the matrices of random coefficients and, naturally, the original data itself. Our results, some of which hold even with finite block lengths, show that, predicated on optimal source-coding, information-theoretic security is achievable for any field size without loss in terms of decoding probability. It follows that protecting the encoding matrix is generally sufficient to ensure confidentiality of network coded data.

The issue of security in network coding is particularly important when it is used in peer-to-peer networking. Using network coding, a peer node can reconstruct the whole file when it has received enough degrees of freedom to decode all the blocks. This scheme is completely distributed, and eliminates the need for a scheduler, as any block transmitted contains partial information of all the blocks that the sender possesses. It has been shown both mathematically and through live trials that the random linear coding scheme significantly reduces the downloading time and improves the robustness of the system.

A major concern for any network coding system is the protection against malicious nodes. Take the above content distribution system for example. If a node in the P2P network behaves maliciously, it can create a polluted block with valid coding coefficients, and then sends it out. Here, coding coefficients refer to the random linear coefficients used to generate this block. If there is no mechanism for a peer to check the integrity of a received block, a receiver of this polluted block would not be able to decode anything for the file at all, even if all the other blocks it has received are valid.

To make things worse, the receiver would mix this polluted block with other blocks and send them out to other peers, and the pollution can quickly propagate to the whole network. This makes coding based content distribution even more vulnerable than the traditional P2P networks, and several attempts were made to address this problem. Some previous work proposed to use homomorphic hash functions in content distribution systems to detect polluted packets, or the suggested the use of a Secure Random Checksum (SRC) which requires less computation than the homomorphic hash function. However, one requires a secure channel to transmit the SRCs to all the nodes in the network. Work from Microsoft proposed a signature scheme based on Weil pairing on elliptic curves and provides authentication of the data in addition to pollution detection, but the computation complexity of this solution is quite high. Moreover, the security offered by elliptic curves that admit Weil pairing is still a topic of debate in the scientific community. In collaboration with Dr. Han (AFRL), Dr. Kaljer (HP Research Labs) and my student, we have developed a homomorphic signature scheme that is not based on elliptic curves, and is designed specifically for random linear coded systems. We view all blocks of the file as vectors, and make use of the fact that all valid vectors transmitted in the network should belong to the subspace spanned by the original set of vectors from the file. Our scheme can be used to easily check the membership of a received vector in the given subspace, and at the same time, it is hard for a node to generate a vector that is not in that subspace but passes the signature test. We have shown that this signature scheme is secure, insofar as it reduces to the Diffe-Hellman problem, and that that the overhead for the scheme is negligible for large files.

The discussion above indicates that there are different approaches to using network coding in security. An important question that emerges is how to guide the choice of approaches. With Professor Barros (University of Porto) and my student, we have studied the transmission overhead associated with three different schemes for detecting Byzantine adversaries at a node using network coding: end-to-end error correction, packet-based Byzantine detection scheme, and generation-based Byzantine detection scheme. In end-to- end error correction, it is known that we can correct up to the min-cut between the source and destinations. However, if we use Byzantine detection schemes, we can detect polluted data, drop them, and therefore, only transmit valid data. For the dropped data, the destinations perform erasure correction, which is computationally lighter than error correction. We show that, with enough attackers present in the network, Byzantine detection schemes may improve the through put of the network since we

choose to forward only reliable information. When the probability of attack is high, a packet-based detection scheme is the most bandwidth efficient; however, when the probability of attack is low, the overhead involved with signing each packet becomes costly, and the generation-based

scheme may be preferred. Finally, we characterize the tradeoff between generation size and overhead of detection in bits as the probability of attack increases in the network, as illustrated in Figure 5 below.

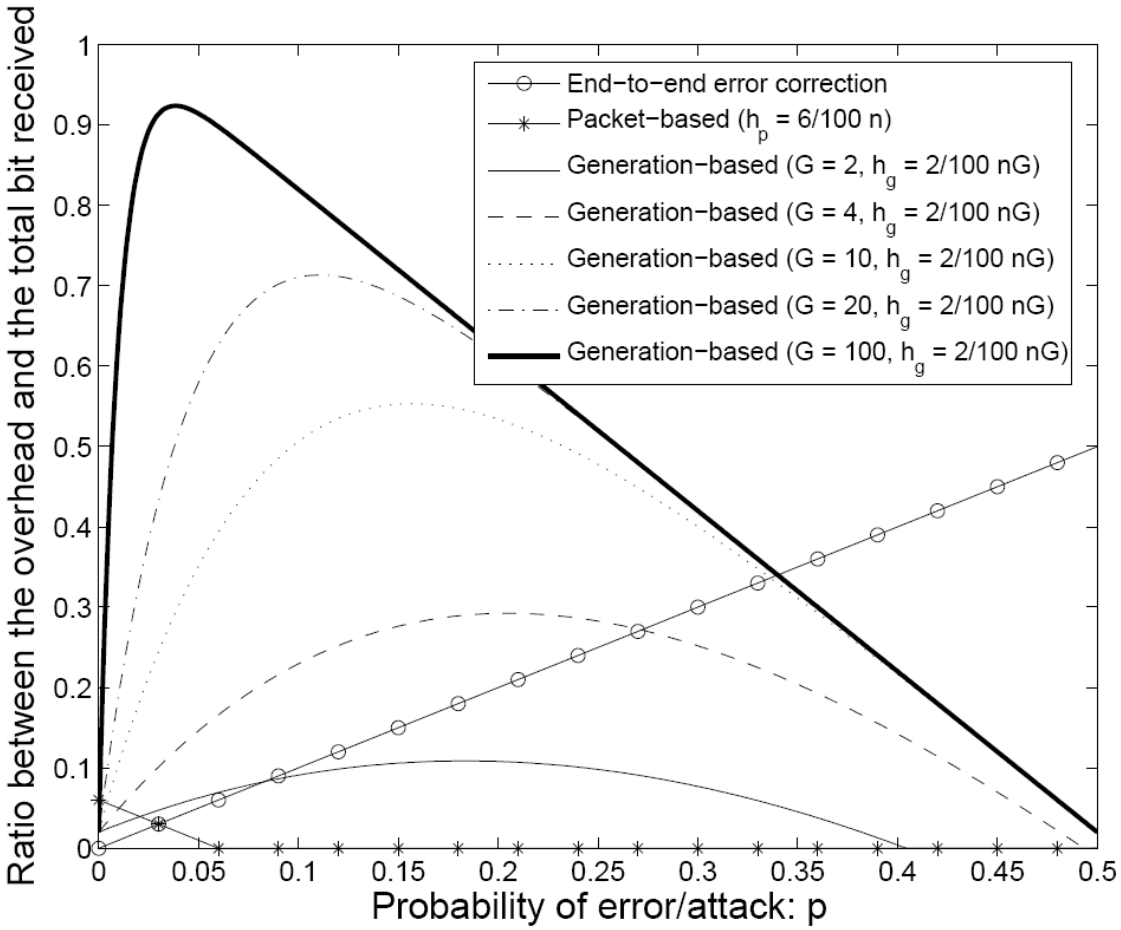


Figure 5: overhead comparison among different approaches to network coding security.

6. Optical networks.

Sponsors:
DARPA FONA and NSF Find

Providing resilient service against failures is a crucial issue for today's optical networks because they operate at very high data rates and thus a single failure may cause a severe loss of data. A variety of protection techniques have been extensively studied for the fault-tolerant operation of optical networks of either ring or mesh topologies. Among them, we particularly focus on the path protection scheme with live back-up, which provides extremely fast recovery, requiring action only from the receiving node. A conventional way to implement such a protection scheme is to transmit two live flows, a primary and a back-up, along link-disjoint paths so that upon link failure the receiver node can switch to the back-up flow. However, it may require an excessive amount of redundant capacity as back-up capacity is not shared among connections. Recent developments have demonstrated that network coding can lead to significant savings in the back-up resources for the multicast scenario protected against link failure by live back-up. An unique

and crucial characteristic of optical networks is converting photonic streams into electronic signals for data processing (O/E/O conversion) is an expensive procedure. Since arbitrary coding operations must be performed in the electronic domain, it appears sensible to restrict the coding

operations only to bitwise XOR, which can be done within the optical domain using a photonic bitwise

7. New directions in Biochemistry.

There are significant intersections between the techniques used in communications and some of the problems that emerge in biochemistry, in particular in genomics, in spectroscopy and inproteolysis. Our group has established new collaborations in this area.

The most widely used chemical experiment for collecting DNA sequencing data is the chain termination method developed by Frederick Sanger in 1977. In the Sanger experiment, a single strand of DNA is mass replicated starting from a fixed primer, but terminated at random locations by fluorescently-labeled markers. The resulting fragments are electrophoretically separated through gels or capillaries by length, which is inversely proportional to their traveling speed. Optical detection for base-specific dyes then gives rise to a set of time series data in the form of a four-component intensity vector, corresponding to the four base-types (Adenine, Cytosine,Guanine,Thymine). The raw data is pre-processed to remove any background intensity level, normalize mobility, and color-correct any correlation between components caused by overlapping dye emission spectra.

Base-calling is the process of identifying the order of DNA bases from pre-processed data, into a sequence of the four base types (A,C,G, T). Owing to random motion of the segments as they pass the detection region, the collected data are successive pulses corresponding to the spread of fragment concentrations around their nominal positions. A typical run, which requires more than 30 minutes to complete, gives approximately six to eight hundred bases, corresponding to 7000 to 10,000 sample points. Given genomes easily contain millions of bases and repetitions are needed to achieve high accuracy in subsequent assembly, a large number of machine days is required to sequence a single genome. In addition, the fixed cost of the machine and variable cost of the reagents sum to thousands of dollars per machine day. To increase the throughput of the overall process while maintaining cost, we mix two DNA segments in electrophoresis, and aim to base call the superposed trace. Figure 1(b) gives a set of sample data, where the average amplitude ratio between the major and minor sequences is close to 2. Here we refer to the sequence with a larger average amplitude as the major, and the other as the minor. To base call a single sequence, an automated sequencer needs to take into account at least three undesirable features of the data: amplitude variation, increasing pulse widths which deteriorate peak resolutions as shown in Figure 1(a), and jitter in peak spacings. Such timing jitter makes it much more difficult to apply a dynamic programming algorithmto resolve the intersymbol interference (ISI), because its inherent random ness makes data association with individual peaks no longer uniform, thus hard to determine.

The most widely used algorithm for base calling a single sequence is Phred, which combines a set of heuristics such as the running average peak spacing, peak areas and concavity measures to determine the bases. Other approaches include parametric deconvolution; combining Kalman prediction of peak locations with dynamic programming to find the maximum likelihood sequence; and performing Markov Chain Monte Carlo methods with a complete statistical model to estimate the peak parameters.

With Doctors Lun and Meldrim (Broad Institute), Professor Koetter (TUM) and our student, we have considered joint base calling of two sequences, as illustrated in Figure 6. A direct extension of these approaches to sequencing two superposed traces is not trivial, for the major and minor traces are not synchronized in time, nor is the separation into the two sequences an easy task. The average amplitude ratio is imperfectly related to the reagent concentration. It can only be set to some range rather than a specific value. We first examine the joint base-calling problem with a

complete statistical model represented graphically on a factor graph. With this setup, the maximum a posteriori (MAP) estimation of individual symbols is very computationally expensive. We propose a two stage model. By viewing the data as similar to pulse amplitude modulated

signals in a communication channel, we first try to find the spike train underlying the mixed sequence data using nonlinear minimum mean square estimation. Next we assign the spikes to the major and minor, to identify the two source sequences. Our MAP approach uses factor graph representations in order to balance local and global effects of base calls.

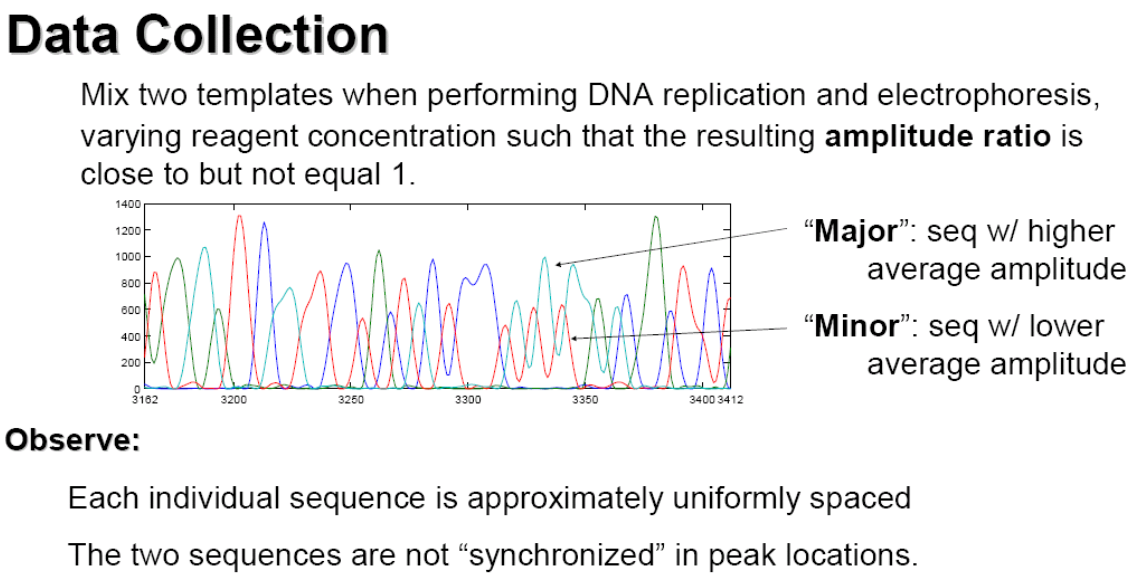


Figure 6: Illustration of mixed sequences.

With Dr. Lun (Broad Institue), Professor Licht and his student, we have considered deconvolution techniques in other application , that of proteolysis. ATP-dependent proteases are processive, meaning that they degrade full-length proteins into small peptide products without releasing large intermediates along the reaction pathway. In the case of the bacterial ATP-dependent protease ClpAP, ATP hydrolysis by the ClpA component has been proposed to be required for processive proteolysis of full-length protein substrates. We present here data showing that in the absence of the ATPase subunit ClpA, the protease subunit ClpP can degrade full-length protein substrates processively, albeit at a greatly reduced rate. Moreover, the size distribution of peptide products from a ClpP-catalyzed digest is remarkably similar to the size distribution of products from a ClpAPcatalyzed digest. The ClpAP- and ClpP-generated peptide product size distributions are fitted well by a sum of multiple underlying Gaussian peaks. Our results are consistent with a mechanism in which ClpP controls product sizes by alternating between translocation in steps of 7-8 ((2-3) amino acid residues and proteolysis. On the structural and molecular level, the step size may be controlled by the spacing between the ClpP active sites, and processivity may be achieved by coupling peptide bond hydrolysis to the binding and release of substrate and products in the protease chamber.

Publications

Journal articles, accepted for publication

1. G. Weichenberg, Chan, V., and Médard, M., "On the Capacity of Optical Networks: A Framework for Comparing Different Transport Architectures," *IEEE Journal on Selected Areas in Communications: Optical Communications and Networking Series*, Volume 25, Issue 6, August 2007, pp. 84 – 101. **
2. J.-K. Sundararajan, Deb, S., and Médard, M., "Extending the Birkhoff-von Neumann Switching Strategy for Multicast-On the use of Optical Splitting in Switches," *IEEE Journal on Selected Areas in Communications: Optical*

Communications and Networking Series, Volume 25 , Issue 6, August 2007, pp. 36-50. **

3. D. S. Lun, Médard, M., Koetter, R., Effros, M., "On coding for reliable communication over packet networks", *Physical Communication*, Volume 1, Issue 1, March 2008, pp. 3-20
4. S. Jaggi, Langberg M., Katti S. , Ho T. , Katabi D., Médard, M., and Effros, E., "Resilient Network Coding in the Presence of Byzantine Adversaries", *Special Issue on Information-theoretic Security of the IEEE Transactions on Information Theory*, Volume 54, Issue 6, June 2008, pp. 2596 - 2603
5. T. Ho, Leong, B., Koetter, R., Médard, M., and Effros, E., "Byzantine Modification Detection in Multicast Networks with Random Network Coding", *Special Issue on Information-theoretic Security of the IEEE Transactions on Information Theory*, Volume 54, Issue 6, June 2008, pp. 2798-2803
6. S. Katti, Hariharan, R., Hu, W. , Katabi, D., and Médard, M., and Crowcroft, J., "XORs in the Air: Practical Wireless Network Coding", *IEEE/ACM Transactions on Networking*, Volume 16, Issue 3, June 2008, pp. 497 - 510
7. S. Jing, Zheng, L., and Médard, M., "On Training with Feedback in Wideband Channels", accepted to the *IEEE Journal on Selected Areas in Communications: Special Issue on Limited Feedback***
8. A. Eryilmaz, Ozdaglar, A., and Médard, M., "On the Delay and Throughput Gains of Coding in Unreliable Networks", accepted to *the IEEE Transactions on Information Theory*
9. D. Lucani, Stojanovic, M., and Médard, M., "Channel Models and Network Coding based Lower Bound to Transmission Power for Multicast" accepted to the *IEEE Journal on Selected Areas in Communications: Special Issue on Underwater Acoustic Networks***
10. P. Youssef-Massaad, Zheng, L., and Médard, M., "Bursty Transmission and Glue Pouring: on Wireless Channels with Overhead Costs", accepted to the *IEEE Transactions on Wireless Communication*

Meeting papers, published

11. L. Lima, Médard, M., and Barros, J., "Random Network Coding: A Free Cypher?", *ISIT* (5 pages), July 2007.
12. M. Kim, Sundararajan, J.-K., and Médard, M., "Network Coding for Speedup in Switches," *ISIT* (5 pages), July 2007. **
13. M. Xiao, Médard, M., and Aulin, T., "A Binary Coding Approach for Combination Networks and General Erasure Networks," *ISIT* (5 pages), July 2007.
14. D. Traskov, Lun, D.S., Koetter, R., and Médard, M., "Network Coding in Wireless Networks with Random Access," *ISIT* (5 pages), July 2007.
15. V. Doshi, Shah, D., and Médard, M., "Source Coding with Distortion through Graph Coloring," *ISIT* (5 pages), July 2007. **
16. F. Zhao, Kalker, T., Médard, M., and Han, K., "Signatures for Content Distribution with Network Coding," *ISIT* (5 pages), July 2007. **
17. S. Katti, Maric, I., Goldsmith, A., Katabi, D., and Médard, M., "Joint Relaying and Network Coding in Wireless Networks," *ISIT* (5 pages), July 2007. **
18. J.-K. Sundararajan, Shah, D., and Médard**,** M., "On Queueing in Coded Networks - Queue Size Follows Degrees of Freedom," **invited** paper, *Information Theory Workshop*, pp. 212-217, July 2007. **
19. F. Zhao, Lun, D., Médard, M., and Ahmed, E., "Decentralized Algorithms for Operating Coded Wireless Networks," **invited** paper, *Information Theory Workshop* (6 pages), September 2007. **
20. D. Lucani, Médard, M., and Stojanovic, M., "Network Coding Schemes for Underwater Networks: The Benefits of Implicit Acknowledgement," *International Workshop on Under Water Networks* (8 pages), September 2007. **

21. K. Han, Ho, T., Koetter, R., Médard, M., and Zhao, F., "On Network Coding for Security," **invited** paper, *MILCOM* (6 pages), October 2007. **
22. E. Ahmed, Eryilmaz, A., Médard, M., and Ozdaglar, A., "On the Scaling Law of Network Coding Gains in Wireless Networks," *MILCOM* (6 pages), October 2007. **
23. M. Kim, Médard, M., Aggarwal, V., and O'Reilly, U.-M., "On the Coding-Link Cost Tradeoff in Network Coding," *MILCOM* (6 pages), October 2007. **
24. I. Maric, Goldsmith, A., and Médard, M., "Information-Theoretic Relaying for Multicast in Wireless Networks", *MILCOM* (6 pages), October 2007.
25. D. Katabi, Fragouli, C., Markopoulou A., Rahul, H., and Médard, M., "Wireless Network Coding: Opportunities and Challenges", *MILCOM* (6 pages), October 2007.
26. S. Katti, Shintre, S., Jaggi, S., Katabi, D., and Médard, M., "Real Network Codes Breaking the All-Or-Nothing Barrier", *45th Allerton Conference on Communication, Control, and Computing,* October 2007
27. A. ParandehGheibi, Eryilmaz, A. Ozdaglar, A., and Médard, M., "Resource Allocation in Multiple Access Channels," **invited** paper, *Asilomar Conference on Signals, Systems and Computers*, November 2007**
28. A. ParandehGheibi, Eryilmaz, A. Ozdaglar, A., and Médard, M., "Dynamic Rate Allocation in Fading Multiple Access Channels," **invited** paper, *ITA,* January 2008**
29. S. Jing, Zheng, L., and Médard, M., "Layered source-channel coding: a distortion-diversity perspective," **invited** paper, *ITA,* January 2008**
30. B.K. Dey, Katti, S., Jaggi, S., Katabi, D., and Médard, M., "'Real' and 'Complex' Network Codes - Promises and Challenges", *Fourth Workshop on Network Coding Theory and Applications (NETCOD),* January 2008, pp. 1-6.
31. A. ParandehGheibi, Eryilmaz, A. Ozdaglar, A., and Médard, M., "Rate Allocation in Fading Multiple Access Channel," *WiOpt,* March-April 2008**
32. X. Shi, Lun, D.S., Meldrim, J, Koetter, R., and Médard, M., "Joint Base-calling of Two DNA Sequences with Factor Graphs", *IEEE International Conferrence on Acoustics, Speech and Signal Processing (ICASSP),* March-April 2008, pp. 2049-2052**
33. D. Lucani, Stojanovic, M., and Médard, M., "On the Relationship between Transmission Power and Capacity of an Underwater Acoustic Communication Channel", *OCEANS 08*, April 2008, paper (071201-059)**
34. J.-K. Sundararajan, Shah, D., and Médard, M., "ARQ for Network Coding", *ISIT 2008*, July 2008**.
35. M. Xiao, Aulin, T., and Médard, M., "Systematic Binary Deterministic Rateless Codes", *ISIT 2008*, July 2008.
36. S. Katti, Katabi, D., Balakrishnan, H., and Médard, M., "Symbol Level Network Coding for Wireless Mesh Networks", Sigcomm 2008, August 2008
37. A. ParandehGheibi, Eryilmaz, A. Ozdaglar, A., and Médard, M., "Information Theory vs. Queueing Theory for Resource Allocation in Multiple Access Channels", **invited** paper, accepted to *PIMRC*, September 2008**
38. D. Lucani, Médard, M., and Stojanovic, M., "A Lower Bound to Transmission Power for Multicast in Underwater Networks using Network Coding", accepted to *ISITA*, December 2008**
39. J.-K. Sundararajan, Shah, D., and Médard, M., "Online network coding for optimal throughput and delay – the three-receiver case", accepted to *ISITA*, December 2008**
40. L. Lima, Barros, J., Vilela, J.-P., and Médard, M., "An Information-Theoretic Cryptanalysis of Randomized Network Coding - is Protecting the Code Enough?", accepted to ISITA, December 2008
41. D. Traskov, Heindlmaier, M., Médard, M., Koetter, R. and Lun, D.S., "Scheduling for Network Coded Multicast: A Conflict Graph Formulation", accepted to the 4th IEEE Workshop on Broadband Wireless Access, December 2008

Meeting paper, presented

September 2007, “Le codage sur réseaux - théorie, applications et nouvelles frontières”, **Plenary Speaker**, GRETSI 2007, Troyes, France

September 2007, “New Directions in Wireless Communications,” **Gilbreth Lecture** to the National Academy of Engineering

January 2008, “On Theory and Practice in Network Coding”, **Keynote Lecture**, Coordinated Science Laboratory at UIUC 3rd annual Student Conference in the areas of Control, Communications and Signal Processing.

February 2008, “Delay and throughput in network coding”, **invited** seminar, Iowa State University

June 2008, “Network Coding”, **invited** course, First School of Information Theory, organized by the IEEE Information Theory Society at Penn State

June 2008, “Network coding and security”, **Keynote Address,** IEEE Workshop on Wireless Network Coding, San Francisco

June 2008, “An Introduction to Network Coding”, short course given at Bretagne Telecom, France

June 2008, “Le codage sur réseaux – principes et applications”, **invited** lecture, at the PRACOM conference, Bretagne Telecom

Theses

Doshi, Vishal, “Functional Compression: Theory and Applications”, February 2008 (co-supervised with Devavrat Shah)

ParandehGhebi, Ali, “Fair Resource Allocation in Multiple AccessChannels”, June 2008 (co-supervised with Asuman Ozdaglar)