On The Capacity Of Wireless Networks: The Relay Case

Michael Gastpar and Martin Vetterli
Communication Systems Department
Ecole Polytechnique Fédérale (EPFL)
Lausanne, Switzerland
Michael.Gastpar@epfl.ch, Martin.Vetterli@epfl.ch

***Abstract*— In [1], Gupta and Kumar determined the capacity of wireless networks under certain assumptions, among them point-to-point coding, which excludes for example multi-access and broadcast codes. In this paper, we consider essentially the same physical model of a wireless network under a different traffic pattern, namely the *relay traffic pattern*, but we allow for arbitrarily complex network coding. In our model, there is only one active source/destination pair, while all other nodes assist this transmission. We show code constructions leading to achievable rates and derive upper bounds from the max-flow min-cut theorem. It is shown that lower and upper bounds meet asymptotically as the number of nodes in the network goes to infinity, thus proving that the capacity of the wireless network with n nodes under the relay traffic pattern behaves like $\log n$ bits per second. This demonstrates also that network coding is essential: under the point-to-point coding assumption considered in [1], the achievable rate is constant, independent of the number of nodes.**

Moreover, the result of this paper has implications and extensions to fading channels and to sensor networks.

I. Introduction

One of the key questions in wireless systems is the capacity of the network, and this under different traffic scenarios, and different constraints (bandwidth, average power, peak power). In the case of networks with base stations, this question is analyzed on a cell by cell basis, by considering the multiple access channel from the mobile users to the base station (uplink), and the broadcast channel from the base station to the users (downlink). This area of research has been very active over the last decades, and is relatively well understood. The case of ad-hoc wireless networks is more recent, and thus less well understood. The additional difficulty stems from the fact that any node can act both as a terminal (sender/receiver of data) and as a *relay* for other transmissions (like, for example, a base station in cell phone networks). Hence, an ad-hoc network has substantially more degrees of freedom than a cell network: any kind of cooperation between the users is permissible. Not surprisingly, these additional features make the determination of capacity much more difficult.

The capacity of multi-terminal systems is a subject studied in multi-user information theory, an area of information theory known for its difficulty, open problems and sometimes counterintuitive results. As a case in point, the separation principle which is a cornerstone result for point-to-point transmission of a source to a destination, does not hold in general for multi-user systems [2, p. 448].

In this context, the question of the capacity of a multi-user mobile system like an ad hoc network is certainly a challenging question. In their landmark paper, Gupta and Kumar [1] gave a formula for the achievable global transmission rate of an adhoc network (in bit-meters per second), and under certain assumptions, showed that one could not achieve a better performance. The key result is that, given n nodes in the unit disk and a uniform traffic pattern, one obtains an aggregate capacity of $O(\sqrt{n})$ bit-meters per second, a somewhat disappointing but not all unexpected result.

A pessimist sees that the rate per user goes to zero as the number of users grows, and an optimist would point out that there are other multiuser scenarios where the total rate is much less (e.g. multi-access, where the sum rate is $O(\log n)$).

The analysis of Gupta and Kumar uses a simple point-to-point coding model. This means that at any given time, a receiver only decodes messages from one sender, considering simultaneous transmissions purely as noise, and similarly, at any given time, a sender transmits information only to one receiver. In that respect, it does not answer the capacity question in an information theoretic sense. In other words, under the same physical constraints, but with a better coding scheme, one could achieve higher rates. Nevertheless, the result presented in [1] certainly points out a basic behavior of current ad hoc networks.

In a recent paper, Grossglauser and Tse [3] modified the model in [1] to include mobility explicitly. Allowing for unbounded delay and using only one-hop relaying (but taking advantage of the mobility), they show a $O(n)$ throughput for a mobile ad hoc network.

In the present paper, we study the capacity of an adhoc network with a very particular traffic pattern, namely a single active link. We call this model the *relay network*, since all nodes (except the sender and receiver nodes) act as relay for the communication. This is schematically rendered in Figure 1. Like in [1], our network is located inside a disk of unit area, sketched by the dashed circle in the figure. The interaction is also identical to [1]: the received signal at some node is the sum of the faded signals from the other nodes plus additive white Gaussian noise. In contrast to [1], two special nodes are selected at random: one is to be the source node, the other the destination node. Those are the two nodes surrounded by the dotted circles. Also in contrast to [1], we do not impose a point-to-point coding model, as described above. Rather, we allow for

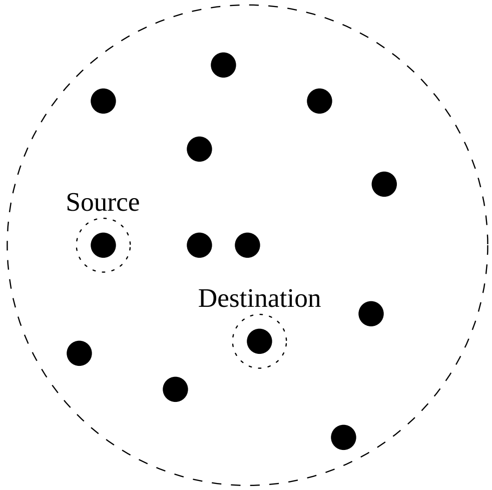

Fig. 1. A wireless network under the relay traffic pattern.

arbitrary cooperation between the nodes, including for example multiple-access and broadcast. In the present paper, we refer to this as "network coding," as opposed to point-to-point coding.

This case, though limited, is amenable to precise analysis, and allows to answer a basic question of adhoc networks, namely what is the precise contribution of relays to the capacity for such a traffic pattern. Interestingly, it is possible to derive upper and lower bounds for the capacity in this case, and the bounds meet as n, the number of nodes in the network, goes to infinity, showing that the capacity is of $O(\log n)$ bits per second. The upper bound follows from the max-flow min-cut theorem as reported in [2, Theorem 14.10.1], which allows for arbitrarily complex network coding. The lower bound follows from a consideration of (almost) uncoded transmission of a particular source across the Gaussian relay network. In this sense, it can be seen as an extension to [4]. Note that a naive (and wrong) use of the throughput result in [1] would give $O(\sqrt{n})$ bits per second, and a more careful application of the point-to-point coding model of [1] yields $O(1)$ only.

The outline of the paper is as follows: In the next section, we recapitulate a few results about multi-user information theory that will in part be used in the sequel. Section III formally introduces the Gaussian relay network, pointing out what is known so far and what an interference model as in [1] would say about its capacity. Section IV studies the capacity of the Gaussian relay network in the limit of a large number of relays, demonstrating the $O(\log n)$ behavior. Finally, Section V discusses the implication of the results and points to open problems.

II. Information Theory And Networks

For an ergodic point-to-point communication problem, information theory provides a set of tools to determine the performance of the best possible coding system. The key ingredient of these tools is that they disregard both delay and complexity, i.e. the code may be infinitely long and infinitely complex if necessary.[1] Under this perspective, information theory permits to determine the best fidelity that one can achieve when a given source has to be transmitted across a given noisy channel. Here, a source is specified by its statistics *and* by a distortion measure. The fidelity is measured with respect to the distortion measure. For the channel, the optimum performance can be characterized by a single number - its capacity. Given this number, one can determine the best achievable fidelity for any source with respect to any distortion measure; no further knowledge of the precise channel structure is required. This is the power of the separation theorem [5, Theorem 21].

Assessing the performance of a network is a trickier issue. Capacity can be generalized to the notion of *capacity region*. For a given statistical description of the network, a set of constraints (power etc.), and a list of desired communications, the capacity region is the closure of all rate tuples that can be achieved simultaneously. A rate tuple specifies the rate for each of the desired communications. It is generally quite difficult to determine and to describe such a capacity region. In the remainder of this section, we give a short portrait of the flavor of the available results. The goal is to illustrate that capacity results for networks are quite limited and often involve certain additional assumptions on the side.

The best-studied case is multiple-access: n terminals communicate to one "base station." To quote just one result, consider the case where the signals of the terminals are simply added together with white Gaussian noise of unit variance, and only this sum signal is observed by the "base station." Suppose that all terminals have the same power P and must transmit at the same rate R. The largest such rate is $R = \log_2(1 + nP)/n$ [2, p. 379].

The broadcast case has also been studied in detail: One "base station" is communicating information to n terminals. The results are less general here. For the broadcast channel, the capacity region is only known when the channel is "degraded." Fortunately, the Gaussian broadcast channel is always degraded, hence its capacity is known [2, p. 380].

Another situation that has been addressed is the relay channel. Suppose that one terminal sends information to another terminal, and in doing so may use the help of a third terminal. Capacity is known for the so-called "physically degraded" relay channel. Under this model, the signal received by the destination node is a degraded version of the signal received by the relay, plus the signal transmitted by the relay. This assumption is somewhat artificial and not always satisfied by real systems; in particular, it is a poor model for the wireless situation. For example, the channel model considered in this paper is not physically degraded. Moreover, capacity is known for relay channels with certain types of noiseless feedback; however, to our knowledge, it is unknown for example for the Gaussian wireless case (i.e. involving noisy feedback between all terminals) [2, p. 430—432].

There is yet a more fundamental limitation to the generalization of capacity to networks. In the (ergodic) point-to-point case, capacity answers all questions, that is, for any source and any distortion measure (by the separation theorem [5, Theorem 21]). In the general network case, there is no such theorem: It is not true that the best communication scheme is achieved by optimally compressing the sources and transmitting the compressed version across the network, using the rates correspond-

[1] The tools have also been modified to apply to the case of finite delay and complexity, but with less success to date.

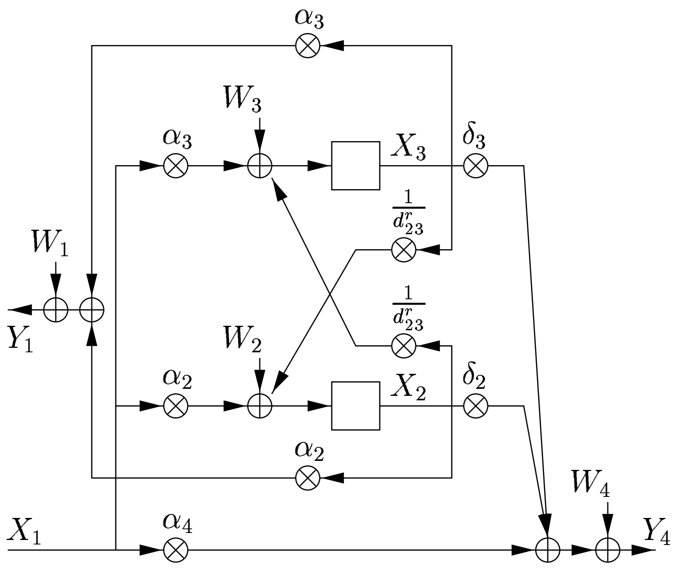

Fig. 2. The Gaussian relay network with two relays, i.e. $n = 4$. The circled cross denotes multiplication with an appropriate fading coefficient, and the circled plus addition of white Gaussian noise. The empty boxes are the two relay decoders/encoders.

ing to a point on the boundary of the capacity region. Examples of this limitation can be found e.g. in [2, p. 448] and in [4, Sec. 5.3].

By this short discussion, we hope to have conveyed the message that the capacity of a wireless network is quite a challenging question in general.

III. The Gaussian Relay Network

A. The Network Model

In this paper, we study the wireless Gaussian relay channel. The underlying physical network could be called the "wireless additive white Gaussian network"; it coincides with the model studied in [1]. In particular, there are n nodes located uniformly in a disk of unit area. During one time slot, each node can only either transmit or receive; it cannot do both simultaneously. The received signal at node k is the linear superposition of the faded transmitted signals from all other nodes and additive white Gaussian noise. This can be written as follows:

$$Y_k \quad = \quad \sum_j \frac{1}{d_{kj}^r} X_j + W_k, \tag{1}$$

where d_{kj} is the Euclidean distance between nodes k and j, r is a positive real number and W_k is additive white Gaussian noise of variance N. (For simplicity, we assume that all noises are of the same variance.) Notice in particular that every node can "hear" every other node. For the case of two relays ($n = 4$), this is illustrated in Figure 2.

Up to here, our model coincides with the model in [1]. The remaining defining elements of our network differ from [1]. At random, one node is selected to be the source, and another node is selected to be the destination. We denote the source node as node 1 and the destination node as node n. The source node can transmit at power $EX_1^2 \leq P$. The $n-2$ nodes that act purely as relays can transmit at a total sum power not exceeding $\sum_{k=2}^{n-1} EX_k^2 \leq (n-2)Q$. Hence, we allow for a certain power allocation between the relay nodes. However, we constrain this power allocation as follows: no single relay may get a power that grows unboundedly with n (the number of nodes in the network).

To simplify notation and since we will use them particularly often, we will denote the fading coefficients from the source node to node k by $\alpha_k = 1/d_{1k}^r$, for $k = 2, \ldots, n$. Similarly, we denote the fading coefficient from any node k to the destination node by $\delta_k = 1/d_{kn}^r$, for $k = 2, \ldots, n-1$.

For the case of only one relay node, this model represents a non-degraded relay channel with noisy feedback. To our knowledge, the capacity of this channel is unknown to date.

B. Previous Results

The channel model described in Section III-A is an extension of the single-relay channel studied in [6]. As mentioned earlier, capacity has been found for the so-called degraded relay channel, and for a certain case of noiseless feedback. Our model (for the case $n = 3$) does not fall into this class. To our knowledge, its capacity is unknown to date.

Another related approach is the one taken by Gupta and Kumar in [1]. They consider the physical network that we described in the Section III-A, i.e. connections between nodes are modeled by Equation (1). The key difference between the consideration in [1] and ours lies in the *traffic pattern*: In [1], all the nodes are split into source/destination pairs uniformly at random. Each source then conveys information exclusively to its assigned destination.

For this situation, [1] strives to determine the maximum throughput, i.e. the maximum number of bit-meters per second that can be transmitted across the network. A solution is found under the additional assumption that all communication is point-to-point. This means that during any given time slot, one node transmits to exactly one other node, and the latter considers all other incoming signals purely as noise, hence excluding any form of network coding (broadcast, multi-access etc.) or decoding (successive cancellation of interference etc.). Under this auxiliary assumption, it is found that the maximum throughput is of the order of $\sqrt{n}$ bit-meters per second, where n is the number of nodes in the network. The "throughput" can be used to answer a number of interesting questions.

First of all, it implies that for a randomly selected source/destination pair, the transmission rate is $1/\sqrt{n}$ bits per second. This is precisely how the throughput is computed in [1].

Then, suppose that the traffic pattern is such that every node wants to speak to its nearest neighbor only. In that case, the communication distance is reduced to $1/\sqrt{n}$. Hence, a throughput of $\sqrt{n}$ bit-meters per second suggests a constant number of bits per second for each source/destination pair. This is indeed the case, as can be verified easily, for example by adapting the proof in [1].

Similarly, consider now the relay traffic pattern: there is only one source/destination pair while the rest of the network is at their service. Can the throughput result be used to determine the maximum rate at which this source/destination pair can communicate? Suppose that source and destination are one meter apart. A naive application of the throughput result would suggest that the rate of transmission for that source/channel pair is $\sqrt{n}$ bits per second, making the throughput again $\sqrt{n}$ bit-meters per second. However, this naive conclusion is incorrect.

In fact, a more careful application of the arguments from the lower bound in [1] to the relay situation yields a constant rate, independent of n.

Clearly, it would be interesting to obtain a result as powerful as that of [1], but without the restriction to point-to-point coding, rather allowing for arbitrarily complex network codes, including for example superposition coding and successive cancellation decoding. One interesting approach in this direction comes again from Gupta and Kumar. In [7], they study the relay case as described in Section III-A, with the difference that they do not allow for power allocation between the relay nodes. Their approach is to consider the set of all possible feed-forward graphs, i.e. the set of all possible forwarding structures from the source to the destination. For each such structure, an achievable rate can be determined. The remaining problem is to optimize over all graphs. However, the latter (combinatorial) problem has no efficient solution to date. Moreover, while this leads to achievable rates, it has not been established in [7] that this approach yields capacity eventually.

C. Outline Of Our Result

The goal of this paper is to determine the capacity for the network model described in Section III-A. More explicitly, this is the maximum rate at which the source node can communicate reliably to the destination node *using arbitrarily complex coding and decoding*. For example, the relays may exchange information with each other in order to coordinate transmission and to reduce interference, or they may use multi-access and broadcast coding techniques to increase the overall efficiency.

In this paper, we determine capacity for the asymptotic case, that is, as the number of relay nodes tends to infinity. To get a capacity expression based on the arguments presented in this paper, we need to add the following two constraints to our network model:

1) Around the source node there is a “dead zone” of nonzero radius; within this zone, there may not be another node. Similarly, there is also a dead zone around the destination node.
2) The source node may only send half of the time.

For this slightly altered network model, we can indeed determine the asymptotic capacity, i.e. we provide an upper and a lower bound on the rates achievable on that channel, and we demonstrate that they coincide as $n \to \infty$.

The upper bound follows from the cut-set bound as it appears in the textbook of Cover and Thomas [2, Theorem 14.10.1]. This bound is sometimes also called “max-flow min-cut,” a short form of saying that the maximum achievable rate is upper bounded by the minimum “cut.” A “cut” is obtained by separating the network into two parts, and evaluating a certain mutual information with respect to this cut. The terminology “max-flow min-cut” actually comes from [8].

The lower bound follows from a somewhat less standard argument. We first explain our argument for the case of a simple (ergodic) point-to-point channel. The channel is defined by a conditional probability density function $p_{Y|X}$, where X is the channel input and Y its output. Moreover, there may be a constraint on the channel input signal X, for example a limitation on the power. To find a lower bound on the capacity of that channel, pick any source, defined by a source probability density function p_S and a distortion measure d. Then, suggest a joint source/channel coding strategy. This strategy has to satisfy all constraints on the channel input signal. The next step is to select any decoding scheme. Clearly, to get good results, the decoder should minimize the overall distortion (under the initially chosen distortion measure d). Once all these elements are fixed, it is a simple matter to determine the resulting average distortion Δ. Then, we have the following statement:

Theorem 1: The capacity of an ergodic channel specified by a conditional probability density function $p_{Y|X}$ and a set of constraints on the channel input signal is at least $C \geq R(\Delta)$, where $R(\cdot)$ denotes the rate-distortion function of some source p_S under some distortion measure d, and Δ is the average distortion (with respect to d) incurred by the transmission of the source p_S across the channel $p_{Y|X}$ using some joint source-channel coding strategy that respects the constraints on the channel input.

Proof: By contradiction: Suppose $C < R(\Delta)$. But then, by the separation theorem, it is not possible to reconstruct the source at fidelity Δ. However, from our joint source-channel code construction, we know that this *is* indeed possible, hence $C \geq R(\Delta)$. ■

Clearly, such a lower bound is particularly interesting when there is a corresponding upper bound, hence the capacity is $C = R(\Delta)$. By considering arbitrarily complex coding and decoding schemes, we can always achieve this case; however, for very complex coding schemes, there is no advantage from using our argument. The advantage of our argument appears in cases where simple joint source-channel coding schemes achieve optimal (or nearly optimal) performance.

For example, we can give a simple lower bound on the capacity of the standard additive white Gaussian noise channel [2, p. 239]. Pick the source to be zero-mean Gaussian with mean-square error distortion. For the sake of the argument, let the source variance be equal to the power constraint on the channel P. Suppose that the encoding is simply uncoded transmission, and the decoding is a scaling by $P/(P+N)$. The achieved distortion for this scheme is found to be $D' = P\sigma^2/(P+\sigma^2)$. Plugging into the rate-distortion function of the Gaussian source of variance P [2, Theorem 13.3.2],

$$R(D) \quad = \quad \begin{cases} \frac{1}{2}\log_2 \frac{P}{D}, & \text{if } 0 \leq D \leq P, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

we find as a lower bound on the capacity of our channel

$$C \quad \geq \quad R(D') = \frac{1}{2}\log_2(1 + P/\sigma^2). \tag{3}$$

With hindsight, this is indeed the capacity of that channel; in other words, for this special case, the lower bounding technique suggested by Theorem 1 gives the best lower bound on capacity by the aid of a very simple coding technique. As a matter of fact, it is well-known that such a simple joint source-channel code achieves optimum performance in the Gaussian-over-Gaussian example. This has been reported e.g. in [9].

Are there other examples where such a simple joint source-channel code achieves optimum performance, and hence $R(\Delta)$ is equal to capacity? This question has been addressed and answered in [4]. As it turns out, there is an infinite supply of such examples.

The key ingredient that makes our argument work is the separation theorem. In general network situations, there is no such statement; however, we will explain below that it holds for the relay network under consideration in this paper.

IV. Capacity Of The Gaussian Relay Network

In this section, we present an upper and a lower bound on the capacity of the considered Gaussian relay network model (as described above in Section III-A) *including* the two additional constraints that were discussed above, namely *(i)*, that there are "dead zones" around the source node and around the destination node, and *(ii)* that the source node may only transmit half of the time.

A. Upper Bound

As mentioned above, our upper bound is derived from the cut-set theorem as it appears in Cover and Thomas [2, Theorem 14.10.1]. For our network, one such bound is the "broadcast cut," i.e. we separate the source node from the rest of the network. This is illustrated in Figure 3.

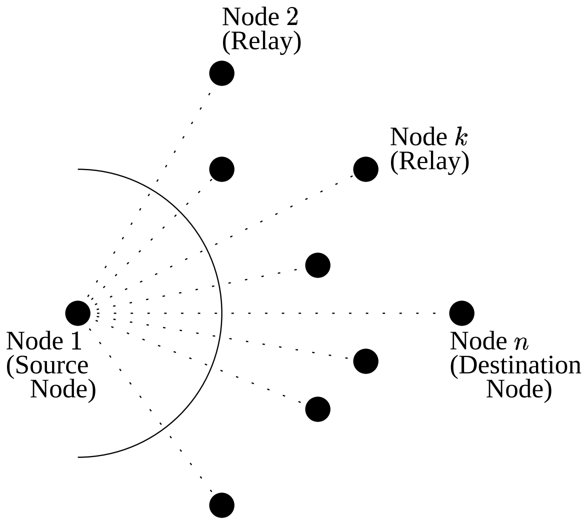

Fig. 3. The "broadcast cut" separates the source node from the rest of the network. The value of the cut-set bound that is used in this paper depends *exclusively* on the dotted connections; any other connection is assumed to be a perfect channel.

The cut-set theorem says that the rate at which we can reliably transmit from the source node to the rest of the relay network cannot exceed the maximum mutual information across this cut, defined as

$$C_{BC} \overset{def}{=} \max I(X_1; Y_2, \ldots, Y_n | X_2, \ldots, X_n). \qquad (4)$$

In particular, to apply Theorem 14.10.1 from [2], we have chosen $S = \{1\}$ and hence $S^c = \{2, \ldots, n\}$ in the notation of the theorem. This maximum can be upper bounded by the capacity of a multiple-antenna channel with one sending antenna and $n-1$ receiving antennas. This problem has been solved in [10] and has become a standard result in information theory. With hindsight, it is quite intuitive that this is an upper bound: We have simply idealized the links between the $n-1$ nodes on the receiving side of the cut as perfect channels, while in reality they are noisy channels. Clearly, this step cannot decrease capacity. But for the system with the idealized channels, we can indeed determine capacity, precisely by using the results from [10]. It can be expressed as

$$C_{BC} = \frac{1}{2}\log_2\left(1 + \frac{||\alpha||^2 P}{N}\right), \qquad (5)$$

where α denotes the vector of length $n-1$ of all the α_i's, i.e. $\alpha = (\alpha_2, \ldots, \alpha_n)$. Consequently, $||\alpha||$ denotes the magnitude of that vector.

Hence, if the source were allowed to transmit in every time slot, this argument would lead to an upper bound of C_{BC}. We will argue later on that this bound must be expected to be loose in general.

However, under the auxiliary constraint that the source node may be transmitting only half of the time, the upper bound on capacity becomes $C \leq \frac{1}{2}C_{BC}$, which establishes the following theorem.

Theorem 2—upper bound: For any particular realization of the random geometry of the network, the capacity of the considered relay network is upper bounded by

$$C \leq C_{upper} \overset{def}{=} \frac{1}{4}\log_2\left(1 + \frac{||\alpha||^2 P}{N}\right). \qquad (6)$$

An upper bound on the *expected* capacity over all possible realizations of the random geometry of the network is then found by taking the expectation of C_{upper} over α. When the nodes are located uniformly at random (as is the case in our network model), $||\alpha||^2$ grows essentially linearly in n. This argument could be made precise; however, under our additional assumption that there is a dead zone around the source node, this point becomes simpler: all the α_k's are bounded above. Therefore, $||\alpha||^2$ cannot grow more than linearly in n. As an intermediate and not very surprising conclusion, this tells us that in any case, capacity behaves at best like $\log n$. Notice that this is a direct consequence of the *traffic pattern*. In other words, in scenarios where the goal is to maximize the sum rate (or total throughput) in the network, it is not a good idea to operate it in pure relay mode: Gupta and Kumar have shown in [1] that a sum rate of $\sqrt{n}$ is achievable if the traffic pattern is comprised of $n/2$ uniformly chosen communicating pairs.

B. Lower Bound

In this section, we use the lower bounding technique that was described above in Theorem 1. To do so, an additional trick is needed. Suppose that the relays operate as follows: If in time slot t, they receive a signal, then they transmit in time slot $t+1$ that exact same signal, scaled to meet their power constraint.

Suppose this system is our new channel. Clearly, this channel cannot have a capacity that is *larger* than the capacity of the Gaussian relay network. At the same time, this new channel is just a simple ergodic point-to-point channel, and hence, the separation theorem does apply. This means that we can indeed use the lower bounding technique that was described above in Theorem 1.

The first step is thus to pick a suitable source. Not surprisingly, for the problem at hand, we select the Gaussian source with squared-error distortion measure.

The second step is the encoding rule. We consider simply uncoded transmission, as follows: In time slot t, the source node broadcasts the source output $X_1(t)$ to all the relays simultaneously and without coding. Consequently, in time slot $t+1$, the relays scale their noisy versions of $X_1(t)$ to their power constraint P_k and forward this to the destination node. In time slot $t+2$, the game starts over with the next source output.

Notice that this strategy satisfies the constraint that the source node may transmit only half of the time.

Finally, for the decoding, the receiver forms the estimate

$$\hat{X}_1(t) \quad = \quad \gamma_1 Y_n(t) + \gamma_2 Y_n(t+1). \qquad (7)$$

The coefficients γ_1 and γ_2 are chosen to minimize the resulting mean-square error which we will denote as D_1.

To compute D_1, we have to determine $Y_n(t)$ and $Y_n(t+1)$. At time t, only the source node is transmitting, and hence

$$Y_n(t) \quad = \quad \alpha_n X_1(t) + W_n(t). \qquad (8)$$

To determine $Y_n(t+1)$ recall that the signal received by relay k at time t is $Y_k(t) = \alpha_k X_1(t) + W_k(t)$. This is scaled to meet the power constraint P_k of relay k. Hence, the signal transmitted by relay k is

$$X_k(t+1) \quad = \quad \sqrt{\frac{P_k}{\alpha_k^2 P + N}}(\alpha_k X_1(t) + W_k(t)) \qquad (9)$$

for $k = 2, \ldots, n-1$. The signal received at the destination node is

$$Y_n(t+1) \quad = \quad \sum_{k=2}^{n-1} \delta_k X_k(t+1) + W_n(t+1). \qquad (10)$$

To simplify notation, we introduce the symbol

$$\beta_k = \delta_k \sqrt{P_k/(\alpha_k^2 P + N)}, \qquad (11)$$

and to make notation more compact, we also introduce the two vectors of significant fading coefficients: $\tilde{\alpha} = (\alpha_2, \ldots, \alpha_{n-1})$ is the vector of fading coefficients from the source node to the $(n-2)$ relays, and $\beta = (\beta_2, \ldots, \beta_{n-1})$ is the vector of fading coefficients from the $(n-2)$ relays to the destination. With this, the mean-square error can be expressed as

$$D_1 \quad = \quad \frac{PN}{\frac{\langle\tilde{\alpha},\beta\rangle^2}{1+||\beta||^2}P + \alpha_n^2 P + N}. \qquad (12)$$

It remains to decide on a favorable power allocation. We choose the (generally suboptimal) allocation that makes $\beta_k = A\alpha_k$, for $k = 2, \ldots, n-1$ and some constant A. That is,

$$P_k \quad = \quad A^2 \frac{\alpha_k^2}{\delta_k^2}(\alpha_k^2 P + N), \qquad (13)$$

where A is chosen to match the sum power constraint $\sum_{k=2}^{n-1} P_k = (n-2)Q$.

Note that at this point, D_1 is completely determined by the involved powers P, Q and N together with the geometry α_k and δ_k, for all k.

The last step is to verify that none of the P_k increases unboundedly with n. This is ensured by the requirement for "dead zones" around the source node and the destination node and by the fact that the network is inside a disk of unit area: both α_k and δ_k are strictly larger than zero and strictly smaller than some constant.

In summary, this leads to the following theorem.

Theorem 3—lower bound: For any particular realization of the random geometry of the network, the capacity of the considered relay network is at least

$$C \quad \geq \quad C_{lower} \quad \overset{def}{=} \quad \frac{1}{4}\log_2 \frac{P}{D_1}, \qquad (14)$$

where D_1 is defined above in Equation (12).

Proof: For the Gaussian source across two uses of the relay network, a distortion of D_1 is feasible. But since the separation theorem applies to this situation, the capacity of two uses of the relay network must be at least $R(D_1)$, where $R(\cdot)$ denotes the rate-distortion function of the Gaussian source with respect to mean-square error distortion. Hence, the capacity of one single use of the relay network must be at least $R(D_1)/2$. Plugging into the rate-distortion function for the Gaussian source [2, Theorem 13.3.2], yields the claimed lower bound. ■

C. Asymptotic Capacity

To obtain a capacity result, it remains to be shown that upper and lower bound coincide. For a finite number n of nodes in the network, this is not true. However, it turns out that asymptotically (as $n \to \infty$), they do coincide. More precisely, the following theorem can be stated.

Theorem 4—asymptotic capacity: The capacity C of the considered relay network is between $C_{lower} \leq C \leq C_{upper}$, where

$$\lim_{n\to\infty} (C_{upper} - C_{lower}) \quad = \quad 0, \qquad (15)$$

for any particular realization of the random geometry of the network. Hence, asymptotically, the capacity of the considered relay network is

$$\frac{1}{4}\log_2\left(1 + \frac{||\alpha||^2 P}{N}\right). \qquad (16)$$

Remark: The convergence established in Theorem 4 depends crucially on the assumption of dead zones around the sender and receiver nodes: the geometry of the network can be arbitrary, but it has to respect the dead zone requirement. This limitation of our result is discussed in detail in Section V-A.

This theorem gives the asymptotic capacity for a *fixed* network geometry, and it directly implies a similar statement about the *expected* capacity over all possible incarnations of the random network, simply by taking expectations over α. We explicitly discuss this issue below.

Proof: To prove this statement, we have to show that the following difference goes to zero:

$$C_{upper} - C_{lower} = \frac{1}{4}\log_2\left(\frac{||\alpha||^2P+N}{N}\frac{D_1}{P}\right). \quad (17)$$

Equivalently, we will show that the expression in parentheses goes to one. In fact, this expression can also be interpreted, as follows: notice that the upper bound on capacity directly implies a lower bound on the distortion achievable for any source with respect to any distortion measure. In particular, we determine this for the case where a Gaussian source is transmitted across *two uses* of the Gaussian relay network. The distortion measure is the mean-squared error. Hence, the lower bound on the average distortion is

$$D_{lower} = D(2C_{upper}) = \frac{PN}{||\alpha||^2P+N}, \quad (18)$$

where $D(\cdot)$ denotes the distortion-rate function of the Gaussian source [2, p. 346].

Our joint source-channel coding scheme also transmits a Gaussian source across two uses of the relay network, and it achieves a distortion of D_1.

Hence, the expression in parentheses in (17) is precisely the ratio D_1/D_{lower}, and the goal of the proof is to show that this ratio tends to one asymptotically.

To this end, the ratio can be written out further as follows:

$$\begin{aligned}\frac{D_1}{D_{lower}} &= \frac{||\alpha||^2P+N}{N}\frac{D_1}{P}\\ &= \frac{(||\tilde{\alpha}||^2+\alpha_n^2)P+N}{\frac{\langle\tilde{\alpha},\beta\rangle^2}{1+||\beta||^2}P+\alpha_n^2P+N}. \quad (19)\end{aligned}$$

Notice that purely for notational convenience, we have replaced $||\alpha||^2$ by $||\tilde{\alpha}||^2+\alpha_n^2$. We choose the power allocation that makes

$$\beta_k = A\alpha_k, \quad (20)$$

for $k = 2, \ldots, n-1$. Under this power allocation, we can simplify

$$\langle\alpha,\beta\rangle = A||\tilde{\alpha}||^2 \quad (21)$$

and

$$||\beta||^2 = A^2||\tilde{\alpha}||^2. \quad (22)$$

This permits to eliminate β from Expression (19) to obtain

$$\frac{D_1}{D_{lower}} = \frac{(||\tilde{\alpha}||^2+\alpha_n^2)P+N}{\frac{(A||\tilde{\alpha}||^2)^2}{1+A^2||\tilde{\alpha}||^2}P+\alpha_n^2P+N}. \quad (23)$$

For our further arguments, we prefer to rewrite this by multiplying both the numerator and the denominator by the term $1+A^2||\tilde{\alpha}||^2$ to obtain

$$\frac{D_1}{D_{lower}} = \frac{(1+A^2||\tilde{\alpha}||^2)\left((||\tilde{\alpha}||^2+\alpha_n^2)P+N\right)}{(A||\tilde{\alpha}||^2)^2P+(1+A^2||\tilde{\alpha}||^2)(\alpha_n^2P+N)}. \quad (24)$$

The constant A has to be determined from the total relay power $(n-2)Q$. Using the power allocation as in Equation (13), A has to satisfy the condition

$$\begin{aligned}(n-2)Q &= \sum_{k=2}^{n-1}P_k\\ &= A^2\sum_{k=2}^{n-1}\frac{\alpha_k^2}{\delta_k^2}(\alpha_k^2P+N) \quad (25)\end{aligned}$$

To simplify the notation and the interpretation of the result, we define the following function:

$$b(n) \stackrel{def}{=} \sum_{k=2}^{n-1}\frac{\alpha_k^2}{\delta_k^2}(\alpha_k^2P+N). \quad (26)$$

Notice that $b(n)$ is a nondecreasing function of n: all terms in the sum are nonnegative. Using $b(n)$, we can express the constant A as

$$A^2 = \frac{(n-2)Q}{b(n)}. \quad (27)$$

This is plugged into Equation (24). Multiplying both the numerator and the denominator by $b(n)$, this permits to express D_1/D_{lower} as

$$\begin{aligned}\frac{D_1}{D_{lower}} = &\frac{(b(n)+(n-2)Q||\tilde{\alpha}||^2)\cdot}{(n-2)Q||\tilde{\alpha}||^4P+}\\ &\frac{\cdot((||\tilde{\alpha}||^2+\alpha_n^2)P+N)}{+(b(n)+(n-2)Q||\tilde{\alpha}||^2)(\alpha_n^2P+N)}. \quad (28)\end{aligned}$$

This rather cumbersome expression can be rewritten in the shape

$$\frac{D_1}{D_{lower}} = \frac{(n-2)||\tilde{\alpha}||^4PQ+\cdots}{(n-2)||\tilde{\alpha}||^4PQ+\cdots} \quad (29)$$

The key step of our proof is to argue that $(n-2)||\tilde{\alpha}||^4$ dominates all other terms that grow with n, both in the numerator and in the denominator. Then, the ratio D_1/D_{lower} indeed tends to one as n grows to infinity.

In the numerator, the only competitor is $b(n)||\tilde{\alpha}||^2$. In the denominator, the competitors are $b(n)$ and $||\tilde{\alpha}||^2$.

It is true under much more general conditions that $(n-2)||\tilde{\alpha}||^4$ dominates these three expressions. However, in the case at hand, we can use again our additional assumptions to simplify the argument.

The assumption of a dead zone around the source node implies that α_k is upper bounded by a constant for all k. Moreover, the fact that the network is located in a disk of unit area implies that δ_k is lower bounded by a constant strictly greater than zero. These two ingredients imply that every term in the sum $b(n)$ as defined in Equation (26) is upper bounded by a constant. Hence, $b(n)$ cannot grow faster than linearly in n.

Moreover, since all terms α_k are strictly larger than zero, it follows that $||\tilde{\alpha}||^2$ is a strictly increasing function of n.

These two insights are sufficient to prove that $(n-2)||\tilde{\alpha}||^4$ indeed dominates both numerator and denominator, as can be verified easily.

To make the argument precise, the limit can then be computed for example by successive applications of the rule of Bernoulli-de l'Hopital. ■

For the proof of Theorem 4, it was not necessary to determine $||\alpha||^2$ precisely. However, to determine capacity, it is still necessary to know how $||\alpha||^2$ behaves as a function of n. We have discussed this right after the upper bound (Theorem 2). There, we argued that under certain conditions, it essentially grows linearly in n. One network structure for which this is true is the case when there are dead zones around the source node and around the destination node. This can be seen by noting that each α_k is upper bounded by a constant.

For the case of a random geometry, it remains to determine the *expected* asymptotic capacity $\overline{C}$ over all possible network realizations,

$$\overline{C} \quad = \quad E_\alpha\left[\frac{1}{4}\log_2\left(1+\frac{||\alpha||^2 P}{N}\right)\right], \qquad (30)$$

where $E_\alpha[\cdot]$ denotes the expectation with respect to the random variable α. Note that the expected asymptotic capacity does not depend on the full geometry of the network; rather, it is sufficiently described by the statistical behavior of the distance from the source node to the relays. For the stochastic model of [1], i.e. for the case where the node locations are selected *uniformly*, it can again be argued that $||\alpha||^2$ grows essentially linearly. While a precise analysis of this case is beyond the scope of the present paper, numerical illustrations of this point will be supplied below.

Notice however that it is possible to construct scenarios where $||\alpha||^2$ grows more than linearly by increasing the relay density very close to the source node as n increases. Our analysis does not apply to such a scenario: in that case, the suggested strategy to prove achievability (Theorem 3) would give unbounded power to a few relays, while many relays would not get any power at all, as follows directly from Equation (13): as the relay node k approaches the source, its corresponding value of α_k grows without bound, and so does its power P_k. We do not consider this a valid (nor an interesting) power allocation. Clearly, a different power allocation strategy may remedy this problem; but this is beyond the framework of the present paper.

D. Numerical Results

Another issue of interest is the behavior of the convergence of upper and lower bounds in Theorem 4 for a random geometry as the number of nodes tends to infinity. In this paper, we do not present a theoretical analysis of this question; rather, we show the result of a numerical simulation.

For the simulation, we generate the network successively. First, the locations of source and destination nodes are chosen, uniformly at random. Then, in each step, the simulation adds one node to the network, uniformly at random, but respecting a dead zone of a certain radius around the source and destination nodes. In each step, the simulation re-evaluates the difference between the rate bounds. Hence, the result of the simulation is a relationship between the number of relay nodes and the difference between the presented upper and lower bounds,

$$C_{upper} - C_{lower} \quad = \quad \frac{1}{2}C_{BC} - \frac{1}{4}\log_2\frac{P}{D_1}, \qquad (31)$$

where D_1 is computed using the power allocation as specified by Equation (13).

Figure 4 shows simulation results for one particular realization of a relay network. For the figure, the parameters have been chosen as follows: the power of the source node is $P = 10$, the average relay power is $Q = 10$ also, and the noise power is $N = 1$. We use a dead zone of radius $0.01R$, where R is the radius of the network, that is, $R = 1/\sqrt{\pi}$ meters. For any given number of relays at the randomly selected locations, the figure shows the discrepancy between upper and lower bound, normalized by the upper bound, i.e. the figure shows the number of nodes versus the quantity

$$\frac{C_{upper} - C_{lower}}{C_{upper}}. \qquad (32)$$

Recall that somewhere in this gap lies the true capacity of that particular relay network. Clearly, a more complete study of the convergence behavior would involve the consideration of the *average* behavior over multiple realizations of the network geometry. However, such a study is beyond the framework of this paper; the goal of this section is merely to illustrate the behavior.

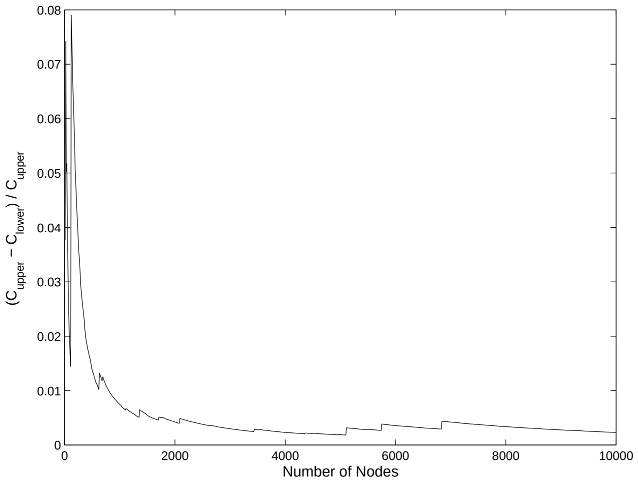

Fig. 4. The normalized difference between the bounds for one realization of an adhoc network.

The reason why the convergence of the two bounds is not unimodal lies to some extent in the randomness of the parameters α_k and δ_k. In particular, the upper bound depends only on α_k while the lower bound depends on both α_k and δ_k. Clearly, if a new relay is added with a high value of α_k (i.e. close to the source node), it considerably increases the upper bound. However, this does not imply that the lower bound increases also:

depending on the corresponding value of δ_k, this will in fact not be the case.

Hence, the fact that the convergence of the bounds is not unimodal is in part also due to our coding scheme, which reacts differently to the node locations than the upper bound. Yet there are cases where the convergence is unimodal, e.g. when $\alpha_k = 1$ and $\delta_k = 1$ for all k.

Recall also that our coding scheme does not make optimum use of the relays, except in the asymptotic case. However, the margin that may (or may not) be gained by a better coding scheme is very small at large n.

V. Discussion And Extensions

A. Discussion

The goal of this paper was to derive the asymptotic capacity of an additive white Gaussian wireless network under a relay traffic pattern. This capacity was found under two additional assumptions.

The first assumption is that there is a dead zone around the source node and another dead zone around the destination node. If this assumption is violated, the proof of Theorem 4 may still work in some cases; however, it invalidates the power allocation. If one of the relay nodes gets very close to the source node, its corresponding value α_k tends to infinity and so does the value of P_k. When there is indeed a relay node arbitrarily close to the source node, a different analysis would have to be performed. For example, under certain conditions, another cut through the network (rather than the one shown in Figure 3) may lead to a tighter upper bound. We could still use the considered joint source-channel coding strategy, but the power allocation would have to be altered. It is not clear whether there is another power allocation under which the strategy performs optimally. For a practical system, the assumption of dead zones does not seem very limiting.

The second assumption is that the source node may only transmit half of the time. Clearly, this assumption is much more restrictive, and it seems "unnecessary." Let us try to give some insight into why it is not easy to obtain a result without this assumption: The "broadcast bound" asks to maximize mutual information across the broadcast cut as illustrated in Figure 3. Clearly, this maximum is achieved when all relays listen to the source. Now suppose that in every time step, all relays only listen to the source. This will give a large value for the upper bound, while in truth very little information is carried through to the destination node. Clearly, the bound should be expected to be loose: some of the relays have to pass the message onwards to the destination node. The assumption that the source node may only transmit half of the time is one way to remedy the weakness of the considered cut-set bound.

There are several other ways that may lead out of this impasse. First, the max-flow min-cut theorem could be used in a more powerful version to take into account the fact that when all relays are listening to the source, then none of them actually forwards messages to the destination. However, it seems that the coding scheme used to prove the lower bound (Theorem 3) would have to be adapted, too, since the source node now sends a new message in every time step. In that case, there is a large interference between what the source node transmits and what the relays transmit, and it is not clear how to handle this case. Another solution could be to study a different kind of nodes: relays that can transmit and receive simultaneously. Our results can also be altered to apply to certain scenarios of this type.

Finally, let us argue that the coding scheme, as simple as it is, is genuine network coding: In the first step (the broadcasting from the source node to the relays), a "code" is used that permits every relay to decode at its particular level of fidelity. This is clearly related to the fact that when one Gaussian source is sent across a Gaussian broadcast channel to multiple destinations, then uncoded transmission is an optimal strategy and actually *outperforms* any approach based on capacity-achieving codes. Extensions of this amazing behavior were presented in [4]. In the second step (the multi-accessing from the relays to the destination), cooperative transmission is used to boost transmit power. It is the combination of these coding steps that yields an achievable rate that behaves like $\log n$. We have already mentioned that if on the contrary, only point-to-point coding is used, then the achievable rate remains constant, independent of n. Hence, for the Gaussian relay network as we have considered it here, network coding significantly changes the asymptotic behavior. This conclusion is certainly of interest in the interpretation of the result of [1]: it suggests the possibility that the asymptotic behavior of capacity *does change* when network coding rather than only point-to-point coding is allowed.

B. Constant Relay Sum Power

In this section, we point to an extension of our results to another case of interest: suppose that the $n-2$ relay nodes have to share a constant power Q. Otherwise, we impose the same conditions like in Section III. It is clear that the upper bound (Theorem 2) is not affected by this change; it still tends to infinity like $\log_2 n$. It seems at first that this upper bound should be much too loose; more precisely, it seems that the capacity should remain finite even though the number of relays tends to infinity since the overall power is finite.

Somewhat surprisingly however, a similar capacity result holds. More precisely, the difference $C_{upper} - C_{lower}$ tends to a constant that is strictly larger than zero (but independent of n). Hence, even in this case, the capacity behaves asymptotically like $\log_2 n$. A more detailed analysis of this case will be presented at a later stage. The reason for this somewhat counterintuitive behavior lies in the amount of spatial diversity provided by our channel model.

C. Limited Network Knowledge

Another interesting feature of the lower bound presented in Theorem 3 of this paper is that it does not require full knowledge of the network. In particular, the relay node k does not need to know the exact locations of the other relay nodes; all it needs in order to determine its appropriate power level P_k is the constant A as in Equation (27). More interestingly, the source node actually need not know anything about the network geometry at all; it simply transmits at its power level.

These issues are of special interest in the presence of *fading*. An analysis of this case will be presented at a later stage.

D. Sensor Networks

The result presented in this paper also implies a result about a certain sensor network situation. Suppose that the underlying phenomenon to be measured is a Gaussian random variable X, and suppose that wireless sensors are scattered around the physical objects such that each of the sensors measures a faded and noisy version of X, where the noise is Gaussian. That is, the measurement of sensor k is

$$Y_k = \alpha_k X + W_k, \tag{33}$$

where W_k is a Gaussian random variable. If a central station that receives the signals from the sensors wants to reconstruct X with respect to the mean-squared error criterion, what coding strategy should the sensors employ? For this case, our result implies that as the number of sensors increases to infinity, it is optimal for the sensors to simply transmit their measurement without any coding at all, using the scheme described in Section IV-B. Clearly, this result can be extended to similar scenarios.

ACKNOWLEDGMENTS

The authors would like to acknowledge helpful discussions with Emre Telatar and Gerhard Kramer.

REFERENCES

[1] P. Gupta and P. R. Kumar, “The capacity of wireless networks,” *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, March 2000.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[3] M. Grossglauser and D. N. Tse, “Mobility can increase the capacity of wireless networks,” in *IEEE INFOCOM 2001, Anchorage, Alaska*, April 2001.

[4] M. Gastpar, B. Rimoldi, and M. Vetterli, “To code, or not to code: On the optimality of symbol-by-symbol communication,” *EPFL DSC Tech. Rep. 026 (submitted to IEEE Trans Info Theory)*, May 2001.

[5] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

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

[7] P. Gupta and P. R. Kumar, “Towards an information theory of large networks: An achievable rate region,” in *IEEE Int. Symp. Info. Theory*, Washington DC, June 2001, (Also submitted to *IEEE Trans Info Theory*, 2001.).

[8] L. R. Ford and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.

[9] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[10] I. E. Telatar, “Capacity of multi-antenna Gaussian channels,” *Bell Labs Technical Memorandum*, June 1995.