Caching in Wireless Networks

Urs Niesen
EECS
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email: uniesen@mit.edu

Devavrat Shah
EECS
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email: devavrat@mit.edu

Gregory Wornell
EECS
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email: gww@mit.edu

Abstract— We consider the problem of delivering content cached in a wireless network of n nodes randomly located on a square of area n. In the most general form, this can be analyzed by considering the $2^n \times n$-dimensional caching capacity region of the wireless network. We propose a communication scheme for transmission of messages cached in the network. This provides an inner bound to the caching capacity region.

I. Introduction

With the continued large-scale deployment of infrastructure, wireless networking remains an area of active research. In this context, unicast and multicast traffic has been widely studied. The influence of caches on the network performance, on the other hand, has received considerably less attention. Nevertheless, the ability to cache data at several places in the network is likely to significantly increase supportable data rates. In this paper, we consider the problem of characterizing achievable rates with caching in wireless networks.

In its most general form, this problem can be formulated as follows. Consider a wireless network with n nodes, and assume a node w in the wireless network requests a message available at the set of caches U (a subset of the n nodes) at a certain rate $\lambda_{U,w}$. The collection of all $\{\lambda_{U,w}\}_{U,w}$ can be represented as a caching traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times n}$. The question is then to characterize the set of achievable caching traffic matrices $\Lambda(n) \subset \mathbb{R}_+^{2^n \times n}$. We will be interested in the behavior of this caching capacity region $\Lambda(n)$ for large wireless networks (i.e., as $n \to \infty$) under random node placement.

A. Related Work

Several aspects of caching in wireless networks have been investigated in prior work. In the computer science literature, the wireless network is usually modelled as a graph induced by the geometry of the node placement. This is tantamount to making a protocol model assumption (as proposed in [1]) about the communication scheme used. The quantity of interest involves the distance from each node to the closest cache that holds the requested message. The problem of optimal cache location for multicasting from a single source has been investigated in [2], [3]. Optimal caching densities under uniform random demand have been considered in [4], [5]. Several cache replacement strategies are proposed, for example, in [6].

To the best of our knowledge, caching has not been directly considered in the information theory literature. However, it can be seen that the problem of optimally transmitting messages held at several caches to a destination is a special case of communicating correlated sources over a noisy network. Indeed, we can consider that each cache has an identical message to send to the same destination. This more general problem of transmitting correlated sources has received considerable attention. Unlike the situation with point-to-point communication, for network communication problems source-channel separation does not hold in general [7]. Hence, the problem of source and channel coding have to be considered jointly. While for some special cases optimal communication strategies for transmitting correlated sources over a noisy network are known (for example, broadcast from a single source with independent network links [8], [9]), the general problem is unsolved.

Finally, a special case of the caching problem considered here, in which each destination has only a single cache (i.e., standard unicast traffic), has been widely studied and is by now well understood. See for example [1], [10]–[19].

B. Our Contribution

We consider the general caching problem from an information theoretic point of view. Compared to the prior work mentioned in the last section, there are several key differences. First, we do not make a protocol channel model assumption, and instead allow the use of arbitrary communication protocols over the wireless network. Second, we allow for general traffic demands, i.e., arbitrary number of caches, and arbitrary demands from each destination. Third, we do not impose that each destination requests the desired message from only the closest cache, nor do we impose that the entire message has to be requested from the same cache. Rather we allow parts of the same message to be requested from distinct caches.

We present a communication scheme for the caching problem, which can be used to transmit general caching traffic. The proposed scheme solves the problems of optimal cache selection, load balancing, and interference and noise mitigation separately, suggesting a layered approach.

C. Organization

The remainder of the paper is organized as follows. In Section II, we introduce the channel model as well as notation.

In Section III, we present the main results of the paper. Section contains IV concluding remarks.

Due to space constraints, we present results here without proofs. They can be found in the journal version of the paper.

II. NETWORK MODEL AND NOTATION

Consider the square $A(n) \triangleq [0, \sqrt{n}]^2$ of area n, and let $V(n) \subset A(n)$ be a set of $|V(n)| = n$ nodes on $A(n)$. We assume the following channel model. The (sampled) received signal at node v and time t is

$$y_v(t) = \sum_{u \in V(n) \setminus \{v\}} h_{u,v}(t) x_u(t) + z_v(t)$$

for all $v \in V(n), t \in \mathbb{N}$, where $\{x_u(t)\}_{u,t}$ are the (sampled) signal sent by the nodes in $V(n)$ at time t. Here $\{z_v(t)\}_{v,t}$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variables with mean 0 and variance 1, and

$$h_{u,v}(t) = r_{u,v}^{-\alpha/2} \exp(\sqrt{-1}\theta_{u,v}(t)),$$

for *path-loss exponent*[1] $\alpha > 2$, and where $r_{u,v}$ is the Euclidean distance between u and v. $\{\theta_{u,v}(t)\}_{u,v}$ is assumed to be i.i.d. with uniform distribution on $[0, 2\pi)$. We either assume that $\{\theta_{u,v}(t)\}_t$ is stationary and ergodic as a function of t which is called *fast fading* in the following, or we assume $\{\theta_{u,v}(t)\}_t$ is constant as a function of t, which is called *slow fading* in the following. In either case, we assume full channel state information is available at all nodes[2], i.e., each node knows all $\{h_{u,v}(t)\}_{u,v}$ at time t. We also impose an average power constraint of 1 on the signal $\{x_u(t)\}_t$ for every node $u \in V(n)$.

Partition $A(n)$ into 4^ℓ squares $\{A_{\ell,i}(n)\}_{i=1}^{4^\ell}$ of sidelength $2^{-\ell}\sqrt{n}$, and let $V_{\ell,i}(n)$ be the nodes in $A_{\ell,i}(n)$. The integer parameter ℓ varies between 0 and $L(n)$ defined as[3]

$$L(n) \triangleq \frac{1}{2}\log(n)\big(1 - \log^{-1/2}(n)\big),$$

The partitions at various levels ℓ form a dyadic decomposition of $A(n)$ as illustrated in Figure 1.

A *caching traffic matrix* is an element $\lambda \in \mathbb{R}_+^{2^n \times n}$. Consider $U \subset V(n)$ and $w \in V(n)$. Assume a message that is requested at destination node w is available at all of the caches U. $\lambda_{U,w}$ denotes then the rate at which node w wants to obtain the message from the caches U. Note that we do not impose here that any particular cache $u \in U$ provides w with the

Fig. 1. Square-grids with $0 \le \ell \le 2$, i.e., with $L(n) = 2$. The grid at level $\ell = 0$ is the area $A(n)$ itself. The grid at level $\ell = 1$ is indicated by dashed lines. The grid at level $\ell = 2$ by dotted lines together. Assume for the sake of example that the subsquares are numbered from left to right and then from bottom to top (the precise order of numbering is immaterial). Then $V_{0,1}(n)$ are all the nodes $V(n)$, $V_{1,1}(n)$ are the nine nodes in the lower left corner (delineated by dashed lines), and $V_{2,1}(n)$ are the three nodes in the lower left corner (delineated by dotted lines).

desired message, rather multiple (or all) of the nodes in U could provide parts of the message. Note also that $\lambda_{U,w}$ and $\lambda_{\widetilde{U},w}$ could both be strictly positive for $U \neq \widetilde{U}$, i.e., the same destination could request more than one message from different collection of caches. We assume that messages for different (U, w) pairs are independent. The *caching capacity region* $\Lambda(n)$ of the wireless network $V(n)$ is the set of all achievable caching traffic matrices $\lambda \in \mathbb{R}_+^{2^n \times n}$.

Example 1. Consider $V(n) = \{u_i\}_{i=1}^4$ with $n = 4$. Assume that u_1 requests a message $m_{\{u_3, u_4\}, u_1}$ available at the caches u_3, and u_4 at rate 1 bit per channel use, and an independent message $m_{\{u_3\}, u_1}$ available only at u_3 at a rate of 2 bits per channel use. Node u_2 requests a message $m_{\{u_3, u_4\}, u_2}$ available at the caches u_3 and u_4 at a rate of 4 bits per channel use. The messages $m_{\{u_3, u_4\}, u_1}, m_{\{u_3\}, u_1}$, and $m_{\{u_3, u_4\}, u_2}$ are assumed to be independent. This traffic pattern can be described by a caching traffic matrix $\lambda \in \mathbb{R}_+^{16 \times 4}$ with $\lambda_{\{u_3, u_4\}, u_1} = 1$, $\lambda_{\{u_3\}, u_1} = 2$, $\lambda_{\{u_3, u_4\}, u_2} = 4$, and $\lambda_{U,w} = 0$ otherwise. Note that in this example node u_1 is destination for two (independent) caching messages, and node u_3 and u_4 serve as caches for more than one message (but these messages are assumed independent). $\diamond$

III. MAIN RESULTS

A. Caching Capacity Region

Let $G = (V_G, E_G)$ be an undirected capacitated graph, constructed as follows. G is a tree with leaf nodes $V(n) \subset V_G$. Leaf nodes in G share the same parent node in G if they fall within the same grid square at level $L(n)$ in $A(n)$. Nodes at level ℓ in the tree G share the same parent node if all their children fall in the same grid square at level $\ell - 1$ in $A(n)$. Note that through this construction, each set $V_{\ell,i}(n)$ for

---

[1]It is worth pointing out that recent results [20] suggest that, under certain assumptions on the location of scattering elements, for $\alpha \in (2, 3)$ and very large values of n, the channel model used here might yield results that are too optimistic. However, the same authors show in [21] that, under different assumptions on the scatterers, the channel model used here is still valid also for $\alpha \in (2, 3)$ and very large values of n. This indicates that the issue of proper channel modelling in the low path-loss regime for very large networks is somewhat delicate and requires further investigation.

[2]The proposed scheme can be shown to work also under weaker assumptions on the availability of CSI. In particular, for $\alpha \ge 3$, no CSI is necessary, and for $\alpha \in (2, 3)$, a 2 bit quantization of the channel state $\{\theta_{u,v}(t)\}_{u,v}$ available at all nodes at time t is sufficient.

[3]Throughout $\log$ and $\ln$ represent the logarithms with respect to base 2 and e, respectively.
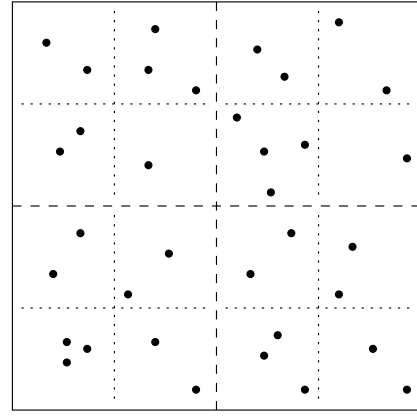
$\ell \in \{0, \dots, L(n)\}$, $i \in 4^\ell$ is represented by exactly one non-terminal node in G. This construction is illustrated in Figure 2. Assign to each edge $e \in E_G$ at level ℓ in G (i.e., between
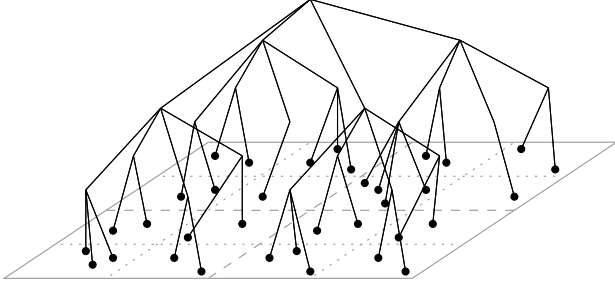

Fig. 2. Construction of the tree graph G. We consider the same nodes as in Figure 1 with $L(n) = 2$. The leaves of G are the nodes $V(n)$ of the wireless network. They are always at level $\ell = L(n) + 1$ (i.e., 3 in this example). At level $0 \le \ell \le L(n)$ in G, there are 4^ℓ nodes. The tree structure is the one induced by the grid decomposition $\{V_{\ell,i}(n)\}_{\ell,i}$ delineated by dashed and dotted lines. Level 0 contains the root node of G.

node levels ℓ and $\ell - 1$) a capacity

$$c_e \triangleq \begin{cases} (4^{-\ell} n)^{2 - \min\{3, \alpha\}/2} & \text{if } 1 \le \ell \le L(n), \\ 1 & \text{if } \ell = L(n) + 1. \end{cases}$$

With slight abuse of notation, we let for $(u, v) = e \in E_G$

$$c_{u,v} \triangleq c_e.$$

As we shall see in the following, a set of messages can be transmitted reliably over the wireless network at rates λ if they can be routed at approximately the same rates between the leaf nodes of G. Given this relation between the wireless network and the graph G, we continue to analyze the caching capacity region (under routing) of G. The caching capacity region of G, in turn, will be shown to be approximately equal to the following quantity:

$$\widehat{\Lambda}(n) \triangleq \left\{ \lambda \in \mathbb{R}_+^{2^n \times n} : \right.$$

$$\left. \sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w} \le \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v} \; \forall S \subset V_G \right\}.$$

The region $\widehat{\Lambda}(n)$ bounds the sum rate for all subsets $S \subset V_G$. Each such subset can hence be understood as a *cut* in the graph G.

We now state the main result of this paper, providing an inner bound on the caching capacity region $\Lambda(n)$.

Theorem 1. *Under either fast or slow fading, for any $\alpha > 2$, there exist $b(n) \ge n^{-o(1)}$ such that*

$$b(n)\widehat{\Lambda}(n) \subset \Lambda(n)$$

with probability $1 - o(1)$ as $n \to \infty$.

B. Computational Aspects

If we consider large wireless networks (i.e., large values of n), computational aspects are of importance. The inner bound to the capacity region $\widehat{\Lambda}(n)$ in Theorem 1 is described in terms of $\Theta(4^n)$ cuts $S \subset V_G$. In other words, computing the entire set $\widehat{\Lambda}(n)$ is not computationally efficient. On the other hand, we shall see that the problem of testing membership of $\lambda \in \widehat{\Lambda}(n)$ can be approximately solved in an efficient manner. More precisely, we show that $\lambda \in \widehat{\Lambda}(n)$ can be checked approximately in polynomial time in the description complexity of λ.

Formally, define for any caching traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}_\lambda(n) \triangleq \sup\{\rho \ge 0 : \rho\lambda \in \widehat{\Lambda}(n)\}.$$

Membership $\lambda \in \widehat{\Lambda}(n)$ can then be evaluated by checking if $\hat{\rho}_\lambda(n) \le 1$. Let $\tilde{\rho}_\lambda(n)$ to be the solution to the following linear program

$$
\begin{array}{lll}
\max & \rho & \\
\text{s.t.} & \sum_{p \in P_{U,w}} f_p \ge \rho\lambda_{U,w} & \forall \, U \subset V(n), w \in V(n), \\
& \sum_{p \in P: e \in p} f_p \le c_e & \forall \, e \in E_G, \\
& f_p \ge 0 & \forall \, p \in P,
\end{array}
\tag{1}
$$

where $P_{U,w}$ is the collection of all paths in G from any node in U to node w (since G is a tree, there are exactly $|U|$ such paths), and

$$P \triangleq \bigcup_{U \subset V(n), w \in V(n)} P_{U,w}.$$

Note that (1), and hence $\tilde{\rho}_\lambda(n)$, can be evaluated in polynomial time in the description length of λ (i.e., in polynomial time in the length of the "input" of the linear program). The next theorem shows that $\tilde{\rho}_\lambda(n)$ is a good approximation to $\hat{\rho}_\lambda(n)$.

Corollary 2. *For any $\alpha > 2$, there exists $b = O(\log(n))$ such that for any caching traffic matrix λ*

$$\tilde{\rho}_\lambda(n) \le \hat{\rho}_\lambda(n) \le b(n)\tilde{\rho}_\lambda(n).$$

As argued above, $\tilde{\rho}_\lambda(n)$ can be computed in polynomial time in the description length of λ. Hence Theorem 2 shows that testing membership $\lambda \in \widehat{\Lambda}(n)$ can be done approximately in polynomial time in the description length of λ.

C. A Content Delivery Protocol

We now describe the communication scheme achieving the inner bound in Theorem 1. The proposed communication scheme consists of three layers, similar to a protocol stack. From high to low level of abstraction, these layers will be denoted by *routing layer*, *cooperation layer*, and *physical layer*. The three layers in this communication scheme serve the following functions. The routing layer performs cache selection and routing, the cooperation layer performs load balancing, and the physical layer deals with interference and noise.

From the view of the routing layer, the wireless network consists of the noiseless capacitated tree graph G defined in Section III-A (see Figure 2 there). To send a message at the

caches U to its destination w, the routing layer routes the message over G. The optimal requests of message parts from the caches in U is found by solving the linear program (1).

The cooperation layer provides the tree abstraction G to the routing layer. Sending a message up or down an edge in the tree G in the routing layer corresponds to distributing or concentrating the same message in the wireless network. More precisely, to send a message from a child node to its parent in G (i.e., towards the root node of G), the message at the wireless nodes in $V(n)$ represented by the child node in G is distributed (over the wireless channel) evenly among all nodes in $V(n)$ represented by the parent node in G. This distribution is performed by splitting the message into equal size parts. To send a message from a parent node to a child node in G (i.e., away from the root node of G), the message at the wireless nodes in $V(n)$ represented by the parent node in G is concentrated on the wireless nodes $V(n)$ represented by the child node in G. This concentration is performed be collecting parts of the previously split up message.

Finally, the physical layer performs this concentration or distribution of messages. Note that the kind of traffic resulting from the operation of the cooperation layer is regular in the sense that within each grid square, all nodes receive data at the same rate. Uniform traffic of this sort is well understood. Depending on the path-loss exponent α, we use either hierarchical relaying [17], [18] (for $\alpha \in (2,3]$) or multi-hop communication (for $\alpha > 3$). It is this operation of each edge in the physical layer that determines the edge capacity of the graph G as seen from the routing layer.

Figure 3 shows an example of the operation of this three layer scheme. For more details on this architecture (in particular the cooperation and physical layer), we refer the reader to [19].

D. Example Scenario

Example 2. (*Nearest Neighbor Cache Selection*)

A reasonable strategy of selecting caches is to request the entire message from the nearest available cache. In fact, this is the strategy implicitly assumed in most of the prior work considering caching in wireless networks cited in Section I-A. This example shows that this strategy can be arbitrarily bad.

Assume $V_{2,1}(n)$ and $V_{2,2}(n)$ are subsets of $V_{1,1}(n)$, and $V_{2,3}(n)$ is a subset of $V_{1,2}(n)$. Consider a node $u^* \in V_{2,2}(n)$, and label the nodes in $V_{2,1}(n) = \{w_1, w_2, \ldots\}$ and in $V_{2,3}(n) = \{u_1, u_2, \ldots\}$. Construct

$$\lambda_{U,w} \triangleq \begin{cases} \rho(n) & \text{if } U = \{u^*, u_i\}, w = w_i \text{ for some } i, \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. This is illustrated in Figure 4.

For every w_i, the nearest cache is u^*, and it can be shown that any communication scheme requesting the entire message from it results in a per node rate of

$$\rho(n) \leq n^{-1+o(1)}.$$

On the other hand, assume each w_i requests the entire message only from the more distant cache u_i. It can be shown that this
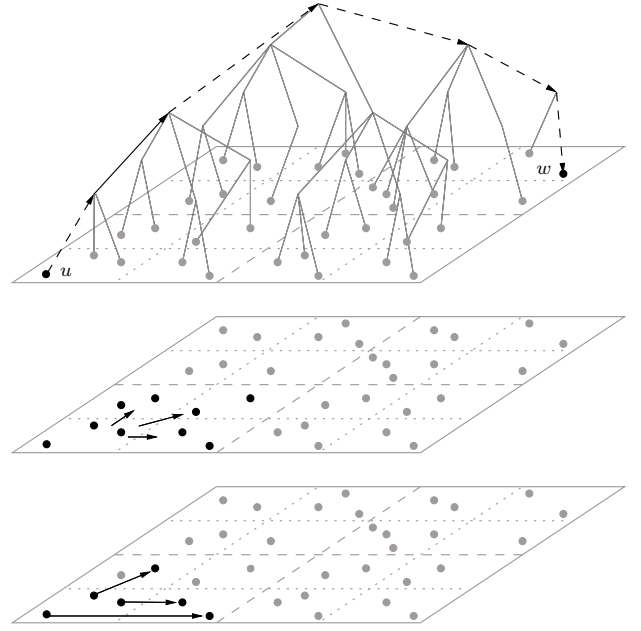

Fig. 3. Example operation of the three layer architecture. The three layers depicted are (from top to bottom in the figure) the routing layer, the cooperation layer, and the physical layer. In this example, we consider a single (U, w) pair. Here, the set of caches U consists of a single node $\{u\}$ in the wireless network shown at the bottom left, and its destination w is in the top right of the network. At the routing layer, the optimal choice of caches is in this case trivial (since there is just one cache u). The optimal route between u and w chosen at the routing layer is indicated in black dashed lines. Consider now the second edge along the path in G from u to w. The middle plane in the figure shows the induced behavior from using this edge in the cooperation layer. The bottom plane in the figure shows (part of) the corresponding actions induced in the physical layer.
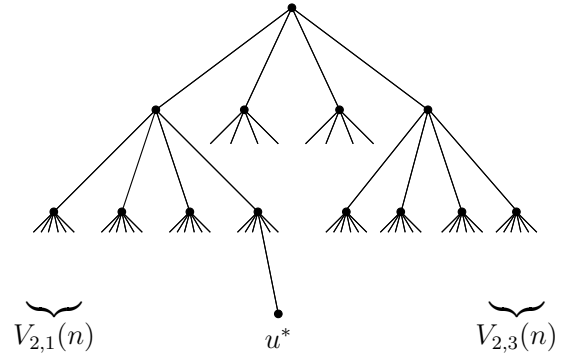

Fig. 4. Caching traffic pattern for Example 2.

is in fact an order-optimal strategy, resulting in a per-node rate of

$$\rho^*(n) = n^{1-\min\{3,\alpha\}/2 - o(1)} \gg n^{-1+o(1)}.$$

Our proposed communication architecture achieves the same order-optimal rate $\rho^*(n)$. This is because the routing layer optimizes over the cache selection instead of simply choosing the closest one. ◇

IV. CONCLUSIONS

We analyzed the influence of caching on the performance of wireless networks. We propose a communication scheme that allows for the transmission of arbitrary caching traffic. This communication scheme consists of three layers, dealing with optimal selection of caches, load balancing, noise and interference, respectively. Using the proposed communication scheme we obtain an inner bound to the caching capacity region. Even though this region, and hence also the inner bound, are $2^n \times n$-dimensional (i.e., exponential in the number of nodes n in the wireless network), we present an algorithm that checks approximate feasibility of a particular caching traffic matrix efficiently (in polynomial time in the description length of the caching traffic matrix).

REFERENCES

[1] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Trans. Inf. Theory*, 46(2):388–404, March 2000.

[2] P. Nuggehalli and C.-F. Chiasserini V. Srinivasan. Energy-efficient caching strategies in ad hoc wireless networks. In *Proc. MobiHoc*, pages 25–34, 2003.

[3] S. Bhattacharya, H. Kim, S. Prabh, and T. Abdelzaher. Energy-conserving data placement and asynchronous multicast in wireless sensor networks. In *Proc. MobiSys*, pages 173–185, 2003.

[4] S. Jin and L. Wang. Content and service replication strategies in multi-hop wireless mesh networks. In *Proc. MSWiM*, pages 79–86, 2005.

[5] B.-J. Ko and D. Rubenstein. Distributed self-stabilizing placement of replicated resources in emerging networks. *IEEE/ACM Trans. Netw.*, 13(3):476–487, June 2005.

[6] L. Yin and G. Cao. Supporting cooperative caching in ad hoc networks. *IEEE Trans. Mobile Comput.*, 5(1):77–89, January 2005.

[7] T. Cover, A. E. Gamal, and M. Salehi. Multiple access channels with arbitrarily correlated sources. *IEEE Trans. Inf. Theory*, 26(6):648–657, November 1980.

[8] T. S. Han. Slepian-Wlf-Cover theorem for network of channels. *Inform. and Control*, 47(1):67–83, January 1980.

[9] J. Barros and S. D. Servetto. Network information flow with correlated sources. *IEEE Trans. Inf. Theory*, 52(1):155–170, January 2006.

[10] L.-L. Xie and P. R. Kumar. A network information theory for wireless communication: Scaling laws and optimal operation. *IEEE Trans. Inf. Theory*, 50(5):748–767, May 2004.

[11] A. Jovičić, P. Viswanath, and S. R. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Trans. Inf. Theory*, 50(11):2555–2565, November 2004.

[12] O. Lévêque and İ. E. Telatar. Information-theoretic upper bounds on the capacity of large extended ad hoc wireless networks. *IEEE Trans. Inf. Theory*, 51(3):858–865, March 2005.

[13] F. Xue, L.-L. Xie, and P. R. Kumar. The transport capacity of wireless networks over fading channels. *IEEE Trans. Inf. Theory*, 51(3):834–847, March 2005.

[14] L.-L. Xie and P. R. Kumar. On the path-loss attenuation regime for positive cost and linear scaling of transport capacity in wireless networks. *IEEE Trans. Inf. Theory*, 52(6):2313–2328, June 2006.

[15] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran. Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Trans. Inf. Theory*, 53(3):1009–1018, March 2007.

[16] P. Gupta and P. R. Kumar. Towards an information theory of large networks: An achievable rate region. *IEEE Trans. Inf. Theory*, 49(8):1877–1894, August 2003.

[17] A. Özgür, O. Lévêque, and D. N. C. Tse. Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Trans. Inf. Theory*, 53(10):3549–3572, October 2007.

[18] U. Niesen, P. Gupta, and D. Shah. On capacity scaling in arbitrary wireless networks. *submitted to IEEE Trans. Inf. Theory*, November 2007. Available online at http://arxiv.org/abs/0711.2745.

[19] U. Niesen, P. Gupta, and D. Shah. The capacity region of large wireless networks. *submitted to IEEE Trans. Inf. Theory*, July 2008. Available online at http://arxiv.org/abs/0809.1344.

[20] M. Franceschetti, M. D. Migliore, and P. Minero. The capacity of wireless networks; information-theoretic and physical limits. In *Proc. Allerton Conference*, September 2007.

[21] M. Franceschetti, M. D. Migliore, and P. Minero. The degrees of freedom of wireless networks: Information theoretic and physical limits. In *Proc. Allerton Conference*, September 2008.