Caching in Wireless Networks

Urs Niesen, *Member, IEEE*, Devavrat Shah, *Member, IEEE*, and Gregory W. Wornell, *Fellow, IEEE*

***Abstract*—We consider the problem of delivering content cached in a wireless network of n nodes randomly located on a square of area n. The network performance is described by the $2^n \times n$-dimensional caching capacity region of the wireless network. We provide an inner bound on this caching capacity region, and, in the high path-loss regime, a matching (in the scaling sense) outer bound. For large path-loss exponent, this provides an information-theoretic scaling characterization of the entire caching capacity region. The proposed communication scheme achieving the inner bound shows that the problems of cache selection and channel coding can be solved separately without loss of order-optimality. On the other hand, our results show that the common architecture of nearest-neighbor cache selection can be arbitrarily bad, implying that cache selection and load balancing need to be performed jointly.**

***Index Terms*—Caching, capacity scaling, multicommodity flow, wireless networks.**

I. Introduction

WIRELESS networks are an attractive communication architecture in many applications as they require only minimal fixed infrastructure. While unicast and multicast traffic in wireless networks has been widely studied, the influence of caches on the network performance has received considerably less attention. Nevertheless, the ability to replicate data at several places in the network is likely to significantly increase supportable rates. In this paper, we consider the problem of characterizing achievable rates with caching in large wireless networks.

In a rather general form, this problem can be formulated as follows. Consider a wireless network with n nodes, and assume a node w in the network requests a message available at the set of caches U (a subset of the nodes) at a certain rate $\lambda_{U,w}$. The collection of all $\lambda_{U,w}$ 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 $\Lambda(n) \subset \mathbb{R}_+^{2^n \times n}$.

Manuscript received August 13, 2009; revised October 15, 2010; accepted October 10, 2011. Date of publication June 25, 2012; date of current version September 11, 2012. This work was supported in part by the Defense Advanced Research Projects Agency under Grant 18870740-37362-C, in part by the Air Force Office of Scientific Research under Grant FA9550-09-1-0317, and in part by the National Science Foundation under Grant CCF-0635191. This paper was presented in part at the 2009 IEEE International Symposium on Information Theory.

U. Niesen was with the Laboratory for Information and Decision Systems and the Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. He is now with the Mathematics of Networks and Communications Research Department, Bell Labs Alcatel-Lucent, New Providence, NJ 07974 USA (e-mail: urs.niesen@alcatel-lucent.com).

D. Shah is with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: devavrat@mit.edu).

G. W. Wornell is with the Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gww@mit.edu).

Communicated by M. Motani, Associate Editor for Communication Networks.

Digital Object Identifier 10.1109/TIT.2012.2205733

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 modeled 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. By definition, such an approach assumes separation of source and channel coding. 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] and [3]. Optimal caching densities under uniform random demand have been considered in [4] and [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, the more general problem of transmitting correlated sources over a network has been studied. Caching is a special case of this problem, in which sources are either independent or identical. While for a single point-to-point channel separation of source and channel coding was shown to be optimal by Shannon [7], the work by Cover *et al.* [8] established that separation is strictly suboptimal for the transmission of correlated sources over a multiple access channel. Hence, even for simple networks, source and channel coding have to be considered jointly. We note that for some special cases separate source and channel coding is optimal, for example, for transmitting arbitrarily correlated sources over a network consisting of independent point-to-point links [9]–[11]. The general problem of joint source-channel coding for noisy networks is unsolved.

Finally, it is worth mentioning the problem of transmitting unicast traffic over a wireless network, which is a special case of the caching problem with each message being available at only a single cache. This problem has been widely studied. Approximate characterizations of the unicast capacity region of large wireless networks (also known as scaling laws) were derived, for example, in [1] and [12]–[22].

B. Summary of Results

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 including joint source-channel

coding. Second, we allow for general traffic demands, i.e., arbitrary number of caches, and arbitrary demands at 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 be requested from the same cache. Rather, we allow parts of the same message to be requested from different caches.

We present a communication scheme for the caching problem, yielding an inner bound on the caching capacity region $\Lambda(n)$. This communication scheme performs separate source and channel coding. For large values of path-loss exponent, we provide a matching (in the scaling sense) outer bound, proving the approximate optimality of our proposed scheme for large values of n. Together, this provides a scaling description of the entire caching capacity region of the wireless network in the large path-loss regime. This result further implies that for caching traffic the loss due to source-channel separation is small (again in the scaling sense) in the large path-loss regime. Since caching traffic is a special case of correlated sources, in which two sources are either identical or independent, this result is a step toward understanding the loss incurred due to source-channel separation for the transmission of arbitrarily correlated sources.

C. Organization

The remainder of this paper is organized as follows. Section II introduces the channel model and notation. Section III presents the main results of the paper. Section IV analyzes the proposed communication scheme and establishes its optimality (up to scaling) for large path-loss exponent. Section V contains concluding remarks.

II. Network Model and Notation

Consider a square of area n, denoted by

$$A(n) \stackrel{\triangle}{=} [0, \sqrt{n}]^2.$$

Let $V(n) \subset A(n)$ be a set of $|V(n)| = n$ nodes placed independently and uniformly at random on $A(n)$. We assume the following complex baseband-equivalent channel model. The received signal at node v and time t is

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

for all $v \in V(n), t \in \mathbb{N}$, and where $x_u[t]$ is the channel input at node u 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

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

for *path-loss exponent* $\alpha > 2$, and where $r_{u,v}$ is the Euclidean distance between u and v. Due to physical constraints, the path-loss exponent α satisfies $\alpha \geq 2$; we adopt the slightly stronger assumption $\alpha > 2$ because it simplifies the statements and derivations of some of the results. The phase terms $(\theta_{u,v}[t])_{u,v}$ are assumed to be i.i.d. with uniform distribution on $[0, 2\pi)$.[1] 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 that $(\theta_{u,v}[t])_t$ is constant as a function of t, which is called *slow fading* in the following. In either case, we assume full channel state information (CSI) is available at all nodes, i.e., each node knows all $(h_{u,v}[t])_{u,v}$ at time t.[2] We also impose an average unit power constraint on the channel inputs $(x_u[t])_t$ for every node $u \in V(n)$.

A *caching traffic matrix* is an element $\lambda \in \mathbb{R}_+^{2^n \times n}$. Consider $w \in V(n)$ and $U \subset V(n)$. 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 requests the message from the caches U.[3] Note that we do not impose that any particular cache $u \in U$ provides w with the desired message, rather multiple nodes in U could provide parts of the message. Note also that $\lambda_{U,w}$ and $\lambda_{\tilde{U},w}$ could both be strictly positive for $U \neq \tilde{U}$, 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 closure of the set of all achievable caching traffic matrices $\lambda \in \mathbb{R}_+^{2^n \times n}$.

Example 1: Consider $V(n) = \{v_i\}_{i=1}^4$ with $n = 4$. Assume that v_1 requests a message $m_{\{v_3,v_4\},v_1}$ available at the caches v_3, and v_4 at rate 1 bit per channel use, and an independent message $m_{\{v_3\},v_1}$ available only at v_3 at a rate of 2 bits per channel use. Node v_2 requests a message $m_{\{v_3,v_4\},v_2}$ available at the caches v_3 and v_4 at a rate of 4 bits per channel use. The messages $m_{\{v_3,v_4\},v_1}$, $m_{\{v_3\},v_1}$, and $m_{\{v_3,v_4\},v_2}$ are assumed to be independent. This traffic pattern can be described by a caching traffic matrix $\lambda \in \mathbb{R}_+^{16 \times 4}$ with $\lambda_{\{v_3,v_4\},v_1} = 1$, $\lambda_{\{v_3\},v_1} = 2$, $\lambda_{\{v_3,v_4\},v_2} = 4$, and $\lambda_{U,w} = 0$ otherwise. Note that in this example node v_1 is destination for two (independent) caching messages, and node v_3 and v_4 serve as caches for more than one message (but these messages are again assumed independent). $\diamond$

To simplify notation, we assume when necessary that large reals are integers and omit $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ operators. For the same reason, we suppress dependence on n within proofs whenever this dependence is clear from the context. We use bold font to denote matrices whenever the matrix structure is of importance. We use the $\dagger$ symbol to denote the conjugate transpose of a matrix. Finally, $\log$ and $\ln$ represent the logarithms with respect to base 2 and e, respectively.

[1] It is worth pointing out that the i.i.d. assumption on the phase terms has to be made with some care. In particular, it is shown in [21], [23], and [24] that this assumption is valid only if the wavelength of the carrier frequency is less than $|A(n)|^{1/2}/n$. For a wide range of scenarios, this is the case, and we assume throughout this paper that this assumption holds.

[2] We make the full CSI assumption in all the converse results in this paper. Achievability can be shown to hold under weaker assumptions on the availability of CSI. In particular, for $\alpha \geq 3$, 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] Note that several rates $\lambda_{U,w}$ are trivial. For example for pairs (U, w) with $w \in U$, or for pairs (U, w) with $U = \emptyset$. We allow these trivial choices for notational convenience. For (U, w) such that $w \in U$, the results will show that $\lambda_{U,w} = \infty$ is achievable; for $U = \emptyset$, they will show that only $\lambda_{U,w} = 0$ is achievable, as would be expected.

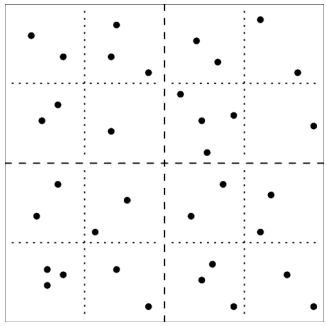

Fig. 1. Subsquares $\{A_{\ell,i}(n)\}$ with $0 \leq \ell \leq 2$, i.e., with $L(n) = 2$. The subsquare at level $\ell = 0$ is the area $A(n)$ itself. The subsquares at level $\ell = 1$ are indicated by dashed lines, the subsquares at level $\ell = 2$ by dotted lines. Assume for the sake of example that 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).

III. Main Results

The main results of this paper are an achievable scheme and an outer bound for the caching capacity region $\Lambda(n)$. Section III-A describes a construction used in Section III-B to establish an inner bound for $\Lambda(n)$. The communication scheme achieving this inner bound respects source-channel separation and is valid for any value of path-loss exponent $\alpha > 2$. In Section III-C, we provide an outer bound that matches (in the scaling sense) the inner bound for large values of path-loss exponent $\alpha > 6$. This leads to an approximate characterization of $\Lambda(n)$ for $\alpha > 6$. This characterization is given in terms of a linear program and is hence computationally tractable as is discussed in Section III-D. The communication architecture achieving the inner bound on the caching capacity region is presented in Section III-E. Various example scenarios are presented in Section III-F.

A. Tree Graph and Linear Program

We describe the construction of a capacitated tree graph induced by the wireless network and a corresponding linear program. These will be needed for the communication scheme achieving the inner bound. This tree graph construction was introduced first in [22].

Partition the square $A(n)$ into 4^ℓ subsquares $\{A_{\ell,i}(n)\}_{i=1}^{4^\ell}$ of side length $2^{-\ell}\sqrt{n}$, and let $V_{\ell,i}(n)$ be the nodes in $A_{\ell,i}(n)$. The integer parameter ℓ varies between 0 and

$$L(n) \stackrel{\triangle}{=} \frac{1}{2}\log(n)\big(1 - \log^{-1/2}(n)\big).$$

The partitions at various levels ℓ form a dyadic decomposition of $A(n)$, as illustrated in Fig. 1. The choice of $L(n)$ is made such that with high probability the number of nodes in each set $V_{L(n),i}$ at the finest grid level is growing to infinity, but not too quickly. See [22] for a detailed discussion.

We now construct an undirected, capacitated tree graph $G = (V_G, E_G)$, as depicted in Fig. 2. The vertex set V_G of G consists of the nodes $V(n)$ in the wireless network plus some additional nodes. The tree G has $L(n)+2$ levels numbered 0 to $L(n)+1$: the root node is at level 0 and leaf nodes are at level $L(n)+1$. The leaf nodes of G are the n nodes $V(n)$ in the wireless network. The nodes of G at level ℓ with $1 \leq \ell \leq L(n)$ are elements of $V_G \setminus V(n)$ and correspond to subsets $\{V_{\ell,i}(n)\}_{i=1}^{4^\ell}$ of the nodes $V(n)$ in the wireless network. The root node of G at level 0 corresponds to all the nodes $V(n)$ in the wireless network. A child node at level $\ell+1$ is connected to a parent node at level ℓ as follows. For $\ell = L(n)$, a node v at level $L(n)+1$ (which is a leaf node of G and hence also an element of the nodes $V(n) \subset V_G$ in the wireless network) is connected to the node in G corresponding to $V_{L(n),i}(n)$ if v belongs to $V_{L(n),i}(n)$. For $0 \leq \ell < L(n)$, a node in G at level $\ell+1$ corresponding to $V_{\ell+1,i}(n)$ is connected to the node in G corresponding to $V_{\ell,j}(n)$ if $V_{\ell+1,i}(n) \subset V_{\ell,j}(n)$.

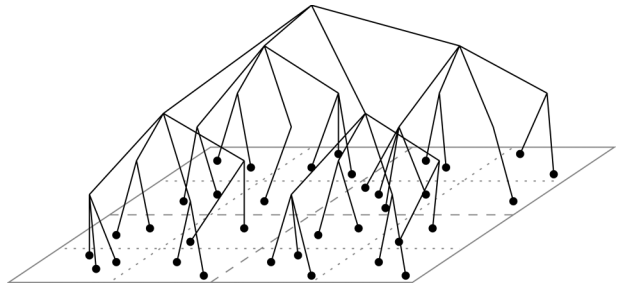

Fig. 2. Construction of the tree graph G. We consider the same nodes as in Fig. 1 with $L(n) = 2$. The leaves of G are the nodes $V(n)$ of the wireless network. They are always at level $\ell = L(n)+1$ (i.e., 3 in this example). At level $0 \leq \ell \leq L(n)$ in G, there are 4^ℓ nodes. The tree structure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

Note that through this construction, each set $V_{\ell,i}(n)$ for $\ell \in \{0,\ldots,L(n)\}$, $i \in 4^\ell$ is represented by exactly one internal node in G. Thus, the cardinality of V_G is

$$\begin{aligned} |V_G| &= |V(n)| + \sum_{\ell=0}^{L(n)} 4^\ell \\ &= n + \frac{1}{3}\Big(4^{L(n)+1} - 1\Big) \\ &\leq 2n. \end{aligned} \tag{1}$$

We assign to each edge $e \in E_G$ at level ℓ in G (i.e., between nodes at levels ℓ and $\ell-1$) a capacity

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

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

$$c_{u,v} \stackrel{\triangle}{=} c_e.$$

The capacity c_e associated with an edge $e = (u,v)$ is to be interpreted as follows. Recall that the nodes u and v in G correspond to a subset of nodes in the wireless network. Let nodes u and v in G be at levels $\ell-1$ and ℓ with $1 \leq \ell \leq L(n)$. The corresponding subsets $V_{\ell-1,i}(n)$ and $V_{\ell,j}(n)$ (for some i and j) have approximately $4^{-\ell+1}n$ and $4^{-\ell}n$ nodes with high probability. Assume we could cooperatively communicate from $V_{\ell-1,i}(n)$ to the nodes $V_{\ell,j}(n)$ in the wireless network. This results in a large multiple-input multiple-output (MIMO) channel with approximately $4^{-\ell+1}n$ transmit and $\frac{3}{4}4^{-\ell}n$ receive antennas. The capacity of this MIMO channel can be evaluated to be approximately $(4^{-\ell}n)^{2-\min\{3,\alpha\}/2}$. Similarly, for a node u at level $L(n)+1$, the capacity from u to the set $V_{L(n),i}$ it is contained

in is approximately equal to one. Thus, we see that the edge capacity c_e is approximately equal to the MIMO capacity between the subsets in the wireless network corresponding to the nodes in G connected by e.

Recall that the leaf nodes of G are equal to the nodes $V(n)$ in of the wireless network. Hence, any caching traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times n}$ for the wireless network is also a valid traffic matrix between leave nodes of G. Assume the leaf nodes of G request messages according to the caching traffic matrix λ. Specifically, we wish to route data from caches in $U \subset V(n)$ to a node $w \in V(n)$ over G at rate $\lambda_{U,w}$. We say that λ is *supportable on* G if this is possible. Let $\Lambda_G(n)$ denote the collection of all caching traffic matrices $\lambda \in \mathbb{R}_+^{2^n \times n}$ that are supportable on G. It can be verified that $\Lambda_G(n)$ is a closed convex set containing the origin.

Given the tree structure of G, there is unique path connecting any two of its nodes. The only way to satisfy the rate demand $\lambda_{U,w}$ by routing is to split it amongst different (u, w) pairs with $u \in U$. Specifically, let $P_{U,w}$ denote the set of $|U|$ unique paths in G between nodes of U and w. For a path $p \in P_{U,w}$ between $u \in U$ and w, let $f_{p,U}$ be the rate at which demand is routed from node $u \in U$ to w along path p for request (U, w). A caching traffic matrix λ is supportable on the capacitated graph G if and only if for each of the $2^n \times n$ pairs (U, w) there exists a decomposition

$$\lambda_{U,w} = \sum_{p \in P_{U,w}} f_{p,U}$$

so that the resulting load on each edge of G is no more than its capacity. Formally, consider the following linear program:

$$\begin{aligned} \max \quad & \phi \\ \text{s.t.} \quad & \sum_{p \in P_{U,w}} f_{p,U} \geq \phi \lambda_{U,w} && \forall U \subset V, w \in V \\ & \sum_{U \subset V} \sum_{w \in V} \sum_{\substack{p \in P_{U,w}: \\ e \in p}} f_{p,U} \leq c_e && \forall e \in E_G \\ & f_{p,U} \geq 0 && \forall U \subset V, w \in V, p \in P_{U,w} \end{aligned} \tag{2}$$

with $V = V(n)$, and where the maximization is over the variables ϕ and $f_{p,U}$. Denote the maximum value of ϕ by $\phi(\lambda)$. The caching traffic matrix λ is supportable on the graph G, if and only if $\phi(\lambda) \geq 1$.

Note that for any $\lambda \in \mathbb{R}_+^{2^n \times n}$, the caching traffic matrix $\phi(\lambda)\lambda$ is supportable on G, i.e., $\phi(\lambda)\lambda \in \Lambda_G(n)$. Thus

$$\phi(\lambda) = \max\Big\{\phi \geq 0 : \phi\lambda \in \Lambda_G(n)\Big\}.$$

In words, $\phi(\lambda)$ is the largest multiple such that the scaled traffic matrix $\phi(\lambda)\lambda$ is supportable on G. Since $\Lambda_G(n)$ is a closed convex set containing the origin, knowledge of $\phi(\lambda)$ for all $\lambda \in \mathbb{R}_+^{2^n \times n}$ completely specifies $\Lambda_G(n)$. We can think of $\phi(\lambda)$, evaluated for all λ, as an equivalent description of the region $\Lambda_G(n)$.

B. Inner Bound

The first result provides an inner bound for the caching capacity region $\Lambda(n)$ in terms of the set $\Lambda_G(n)$ of supportable caching traffic matrices over the graph G. This result is valid for all $\alpha > 2$, i.e., for all values of the path-loss exponent α of interest (excluding the boundary point $\alpha = 2$ as discussed in Section II).

For $\lambda \in \mathbb{R}_+^{2^n \times n}$, define

$$\rho(\lambda) \stackrel{\triangle}{=} \max\Big\{\rho \geq 0 : \rho\lambda \in \Lambda(n)\Big\}.$$

In words, $\rho(\lambda)$ is the largest multiple such that the scaled traffic matrix $\rho(\lambda)\lambda$ is achievable over the wireless network. The caching capacity region $\Lambda(n)$ is a closed convex set containing the origin, and hence, $\rho(\lambda)$ is an equivalent description of $\Lambda(n)$.

Theorem 1: Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1(n) \geq n^{-o(1)}$ such that

$$\rho(\lambda) \geq b_1(n)\phi(\lambda)$$

for all $\lambda \in \mathbb{R}_+^{2^n \times n}$ with probability $1 - o(1)$ as $n \to \infty$.

The proof of Theorem 1 is provided in Section IV-A. We point out that Theorem 1 holds only with probability $1 - o(1)$ for different reasons in the fast and slow fading case. For fast fading, the theorem holds only for node placements that are "regular" enough. A random node placement satisfies these regularity conditions with high probability as $n \to \infty$. For slow fading, Theorem 1 holds under the same regularity conditions on the node placement, but additionally only holds with probability $1 - o(1)$ for the realization of the channel gains.

Given the equivalence of $\rho(\lambda), \phi(\lambda)$ and $\Lambda(n), \Lambda_G(n)$ as mentioned above, Theorem 1 states that $b_1(n)\Lambda_G(n) \subset \Lambda(n)$ with high probability. This links the tree graph G to the wireless network: Every caching traffic matrix that can be routed over the graph G can also (up to a small, in the scaling sense, factor) be transmitted reliably over the wireless network.

The communication scheme achieving the inner bound in Theorem 1 consists of three layers. The lower two layers handle channel coding and load balancing, and effectively transform the wireless network into the tree graph G. The top layer assigns caches to destination nodes and routes data over G. Thus, this scheme performs separate source coding (in the top layer) and channel coding (in the two bottom layers). See Section III-E for a detailed description of this communication architecture.

C. Outer Bound

The next result provides an outer bound for the caching capacity region $\Lambda(n)$ in terms of the $\Lambda_G(n)$. This result is valid for $\alpha > 6$, i.e., for large path-loss exponents.

Theorem 2: Under either fast or slow fading, for any $\alpha > 6$, there exists $b_2(n) \leq n^{o(1)}$ such that

$$\rho(\lambda) \leq b_2(n)\phi(\lambda)$$

for all $\lambda \in \mathbb{R}_+^{2^n \times n}$ with probability $1 - o(1)$ as $n \to \infty$.

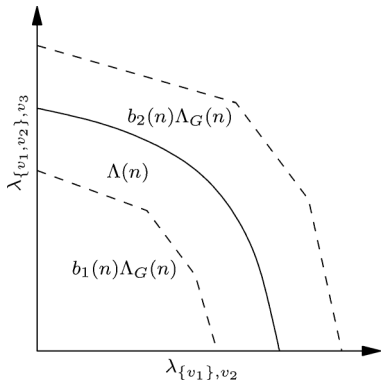

Fig. 3. For $\alpha > 6$, the set $\Lambda_G(n)$ approximates the caching capacity region $\Lambda(n)$ of the wireless network in the sense that $b_1(n)\Lambda_G(n)$ (with $b_1(n) \geq n^{-o(1)}$) provides an inner bound to $\Lambda(n)$ and $b_2(n)\Lambda_G(n)$ (with $b_2(n) \leq n^{o(1)}$) provides an outer bound to $\Lambda(n)$. The figure shows two dimensions (namely $\lambda_{\{v_1\},v_2}$ and $\lambda_{\{v_1,v_2\},v_3}$) of the $2^n \times n$-dimensional sets $\Lambda(n)$ and $\Lambda_G(n)$.

The proof of Theorem 2 is provided in Section IV-B. As with Theorem 1, Theorem 2 holds with probability $1 - o(1)$ for the realization of the node placement and, in the slow fading case, the realization of the channel gains.

Using again the equivalence of $\rho(\lambda)$, $\phi(\lambda)$ and $\Lambda(n)$, $\Lambda_G(n)$, Theorem 2 states that $\Lambda(n) \subset b_2(n)\Lambda_G(n)$ with high probability. Comparing Theorems 1 and 2, we see that, for $\alpha > 6$ and with high probability

$$n^{-o(1)}\phi(\lambda) \leq \rho(\lambda) \leq n^{o(1)}\phi(\lambda)$$

for all $\lambda \in \mathbb{R}_+^{2^n \times n}$ or, equivalently

$$n^{-o(1)}\Lambda_G(n) \subset \Lambda(n) \subset n^{o(1)}\Lambda_G(n).$$

In other words, for $\alpha > 6$, the set of caching traffic matrices $\Lambda_G(n)$ supportable by routing over the tree graph G scales as the caching capacity region $\Lambda(n)$. This is illustrated in Fig. 3.

D. Computational Aspects

Theorems 1 and 2 show that, for large α, $\Lambda_G(n) \approx \Lambda(n)$. Computationally, the question of interest is that of membership, i.e., determining if a given $\lambda \in \mathbb{R}_+^{2^n \times n}$ belongs to $\Lambda(n)$ or, equivalently, determining if $\rho(\lambda) \geq 1$. Since $\rho(\lambda) \approx \phi(\lambda)$, computation of $\phi(\lambda)$ answers the membership question approximately (up to a multiplicative error of $n^{o(1)}$).

The linear program (2) defining $\phi(\lambda)$ can be solved in polynomial time in the number of its constraints and variables [25]. Define

$$\|\lambda\|_0 \stackrel{\triangle}{=} |\{(U,w) : \lambda_{U,w} > 0\}|$$

as the number of (U,w) pairs with positive demand $\lambda_{U,w} > 0$. The number of constraints in the linear program (2) scales linearly in $|E_G| + \|\lambda\|_0$. And the number of variables scales as $n\|\lambda\|_0$. Noting that $|E_G|$ is polynomial in n by (1), this implies that the approximate membership of any λ in $\Lambda(n)$ can be checked in time polynomial in n and $\|\lambda\|_0$.

Note that this need not be polynomial in n, since $\|\lambda\|_0$ could be exponential in n. However, even just to ask the membership query, one needs to specify $\|\lambda\|_0$ distinct numbers. Therefore, the above discussion shows that the computational cost of approximate membership testing takes time polynomial in the effective problem statement, which is the best one can hope for. Moreover, in many situations of practical interest, the number of pairs (U,w) with positive demand can be expected to be only polynomial in the network size n. In these cases, approximate membership can be tested in polynomial time also in n.

E. Content Delivery Protocol

Theorem 1 provides an inner bound for the caching capacity region of a wireless network. We now describe the communication scheme achieving this inner bound. The matching outer bound shows that, for $\alpha > 6$, this scheme is optimal in the scaling sense.

Our proposed communication scheme consists of three layers, similar to a protocol stack. From the highest to lowest level of abstraction, these three layers are the *data layer*, the *cooperation layer*, and the *physical layer*. From the view of the data layer, the wireless network is treated as the abstract capacitated tree graph G, up to a loss of a factor $b_1(n)$ in the capacity of each link. Let us assume that $\frac{1}{b_1(n)}\lambda \in \Lambda_G(n)$. Solve the corresponding linear program (2), and let $f = (f_{p,U})$ be its solution. Since $\frac{1}{b_1(n)}\lambda \in \Lambda_G(n)$, routing traffic according to this solution f allows to support the caching traffic matrix λ in this layer. The next two layers transform this routing solution f for λ over the graph G into a communication strategy for the wireless network.

The cooperation layer provides this tree graph abstraction to the data layer. Recall that the leaf nodes of G are the nodes $V(n)$ of the wireless network and that each internal node of G represents a subset of nodes $V_{\ell,i}(n) \subset V(n)$ within the subsquare $A_{\ell,i}(n)$ in the wireless network. The cooperation layer provides the tree abstraction G by ensuring that, whenever a message is located in the data layer at a particular node v, the message is evenly distributed in the wireless network among the nodes $V_{\ell,i}(n)$ represented by the node v. Recall that the sets $\{V_{\ell,i}(n)\}$ are nested and increasing as ℓ decreases. Hence, as a message travels toward the root node in G in the data layer, it is distributed over a larger area in the wireless network by the cooperation layer. Similarly, as a message travels away from the root node in G in the data layer, it is concentrated on a smaller area in the wireless network by the cooperation layer. Thus, sending a message up or down an edge in the tree G in the data layer corresponds in the cooperation layer to distributing or concentrating the same message in the wireless network (see also Fig. 4).

Formally, this distribution and concentrating of messages is performed as follows. To send a message from a child node to its parent in G (i.e., toward the root node of G), the message at the wireless nodes in $V(n)$ represented by the child node in G is evenly distributed over the wireless channel among all nodes in $V(n)$ represented by the parent node in G. This distribution is performed by splitting the message at each node in $V(n)$ represented by the child node in G into equal sized parts and by transmitting one part to each node in $V(n)$ represented by the parent node in G. 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

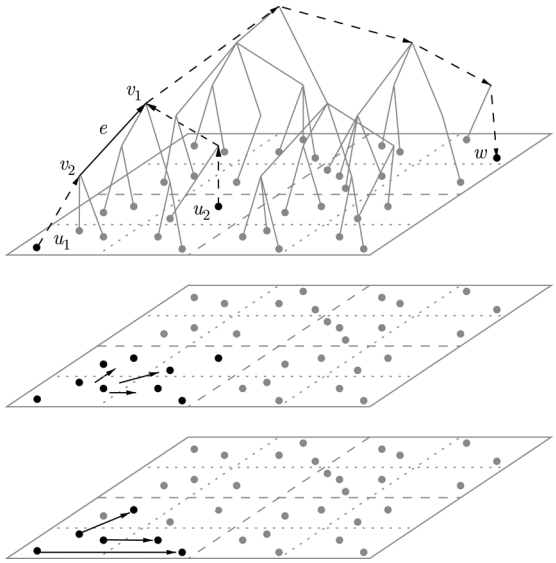

Fig. 4. Example operation of the three-layer architecture. A message available at the caches $U = \{u_1, u_2\}$ is requested at the destination node w. The figure shows the induced actions by this request in the data layer (top plane), cooperation layer (middle plane), and physical layer (bottom plane).

node in G is concentrated on the wireless nodes in $V(n)$ represented by the child node in G. This concentration is performed be collecting at each node in $V(n)$ corresponding to the child node in G the message parts of the previously split up message located at the nodes in $V(n)$ corresponding to the parent node in G.

Finally, the physical layer performs this concentration and distribution of messages induced by the cooperation layer over the physical wireless channel. Note that the kind of traffic resulting from the operation of the distribution or cooperation is highly uniform in the sense that within each subsquare, 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 cooperation [19], [20] (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 data layer.

Note that the value of the path-loss exponent α only significantly affects the operation of the physical layer. The cooperation layer is completely invariant under changes in α, and the data layer is only affected through the value of the edge capacities of the graph G. In particular, even when $\alpha > 3$ so that the physical layer performs multihop communication, the construction of the tree structure G is still necessary. In fact, the role of routing over G can be understood as load balancing of traffic, which is required no matter how the physical layer operates.

We point out that this scheme respects source-channel separation. In fact, source coding is only performed at the data layer (through the selection of message parts from the various available caches). Channel coding is only performed in the cooperation and physical layers.

The next example illustrates the operation of this scheme. For more details on this architecture, see [22].

Example 2: Consider the three layers of the proposed communication architecture depicted in Fig. 4. From top to bottom in the figure, these are the data layer, the cooperation layer, and the physical layer. In this example, we consider a single (U, w) pair. The set of caches U consists of two nodes $\{u_1, u_2\}$ in the wireless network shown at the bottom left, and the corresponding destination w is in the top right of the network. At the data layer, traffic is balanced by choosing which fraction of the message requested at w and available at U is delivered from each node u_1 and u_2 in U. This load balancing is performed by solving the linear program (2). In this simple example, a reasonable choice is to deliver half the message from u_1 and half from u_2. The routes between $\{u_1, u_2\}$ and w chosen at the data layer are indicated in black-dashed lines.

Consider now the second edge along the path in G from u_1 to w labeled by $e = (v_2, v_1)$ in the figure. The middle plane in the figure shows the induced behavior in the cooperation layer from using this edge in the data layer. Note that v_2 and v_1 are not leaf nodes of G, and hence correspond to subsets of $V(n)$ through the construction of G. Let $V_{2,i}(n)$ and $V_{1,j}(n)$ be the subsets of $V(n)$ corresponding to v_2, and v_1, respectively. Since v_2 is a child node of v_1, we must have $V_{2,j} \subset V_{1,i}$. When a message is present at v_2 in the data layer, it is distributed evenly over the three nodes in $V_{2,i}(n)$ in the cooperation layer; in other words, each of the three nodes in $V_{2,i}(n)$ has access to a distinct third of the original message. To send the message over edge e from v_2 to v_1 in the data layer, the cooperation layer splits the message part at each node in $V_{2,i}(n)$ into smaller parts and distributes these subparts evenly over the nodes in $V_{1,j}(n)$. Thus, when the message reaches v_1 in the data layer, each of the nine nodes in $V_{1,j}(n)$ has access to a distinct ninth of the original message in the cooperation layer.

The bottom plane in the figure shows part of the corresponding actions induced in the physical layer. The distribution of message parts from $V_{2,i}(n)$ to $V_{1,j}(n)$ is properly scheduled to minimize interference, and channel coding is performed. The precise nature of the operation of this layer depends on the path-loss exponent α, as explained previously. $\diamond$

F. Example Scenarios

We provide two examples illustrating various aspects of the caching capacity region. Example 3 shows that the strategy of always selecting the nearest cache can be arbitrarily bad. Example 4 illustrates the potential benefit of caching on achievable rates in the wireless network.

Example 3 (Nearest-Neighbor Cache Selection): A simple and intuitive strategy for 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 on caching in wireless networks cited in Section I-A. This example shows that this strategy can be arbitrarily bad.

We consider the scenario illustrated in Fig. 5. Assume $V_{2,1}(n)$ and $V_{2,3}(n)$ are subsets of $V_{1,1}(n)$, and $V_{2,16}(n)$ is a subset of $V_{1,4}(n)$. Consider a node $u^\star \in V_{2,3}(n)$ geographically close to $V_{2,1}(n)$, and label the nodes in $V_{2,1}(n) = \{w_1, w_2, \ldots\}$ and in $V_{2,16}(n) = \{u_1, u_2, \ldots\}$. Construct the traffic matrix

$$\lambda_{U,w} \triangleq \begin{cases} 1, & \text{if } U = \{u^\star, u_i\}, w = w_i \text{ for some } i \\ 0, & \text{otherwise.} \end{cases}$$

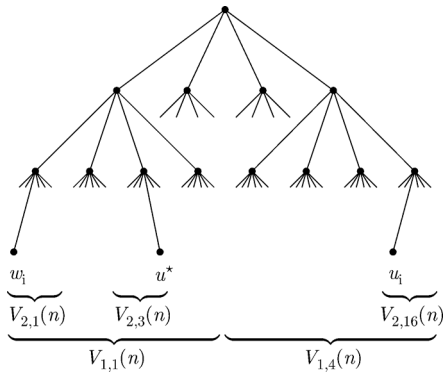

Fig. 5. Caching traffic pattern for Example 3. Each destination node $w_i \in V_{2,1}(n)$ requests a message available at a dedicated cache $u_i \in V_{2,16}(n)$ and at a shared cache $u^\star \in V_{2,3}(n)$.

In words, each node in $w_i \in V_{2,1}(n)$ requests a message available at a dedicated cache $u_i \in V_{2,16}$ and at a shared cache $u^\star \in V_{2,3}$. We want to determine $\rho(\lambda)$, the largest multiple of λ such that the resulting traffic matrix is achievable in the wireless network. In this setting with unit demands, $\rho(\lambda)$ can also be interpreted as the largest uniformly achievable per-node rate.

For every destination node w_i, the nearest cache (both in terms of geographic as well as graph distance) is $u^\star$. Assume each node w_i requests the entire message from its nearest cache $u^\star$. It is easy to show that each node in the wireless network, and, in particular, node $u^\star$, can reliably transmit information at a sum rate of at most $n^{o(1)}$. With high probability, there will be $\Theta(n)$ nodes in $V_{2,1}(n)$ requesting a message at equal rate from $u^\star$. Hence, this strategy achieves a per-node rate of at most $n^{-1+o(1)}$ regardless of the value of the path-loss exponent $\alpha > 2$.

Assume now each w_i uses only the more distant cache u_i. The routes from u_i to w_i for different values of i intersect only at the four edges closest to the root node of G. These four edges have a capacity of order $\Theta(n^{2-\min\{3,\alpha\}/2})$, and hence it can be seen that over the graph G these messages can be routed at a per-node rate of $\Theta(n^{1-\min\{3,\alpha\}/2})$. Together with Theorem 1, this shows that

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

is achievable in the wireless network with high probability. For this simple example, it is easily checked that this strategy is order-optimal for routing over the graph G. Together with Theorem 2, this confirms that, for $\alpha > 6$, no scheme can achieve a better scaling in the wireless network. Hence

$$\rho(\lambda) = n^{1-\min\{3,\alpha\}/2\pm o(1)}$$

for $\alpha > 6$.[4] With some additional work, it can be shown that this is the correct scaling of $\rho(\lambda)$ also for $\alpha \in (2,6]$. This shows that the strategy of always selecting the nearest cache can result in a scaling exponent that is considerably worse than what is achievable with optimal cache selection. $\diamond$

[4]The notation $n^{\pm o(1)}$ is used to indicate that $n^{o(1)}$ is an upper and $n^{-o(1)}$ is a lower bound.

Example 4 (Complete Caches): Assume we randomly pick n^β caches for $\beta \in (0,1)$, each holding a complete copy of all the messages. More precisely, letting $U^\star = \{u_i\}_{i=1}^{n^\beta}$ be the collection of caches, we consider a caching traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times n}$ of the form

$$\lambda_{U,w} = \begin{cases} 1, & \text{if } U = U^\star \\ 0, & \text{otherwise} \end{cases}$$

for every (U,w). In other words, every node $w \in V(n)$ requests a message that is available at a common set of caches $U^\star$. As before, $\rho(\lambda)$ can in this setting with uniform demands be interpreted as the largest uniformly achievable per-node rate.

Assume every node chooses the nearest cache (as discussed in Example 3). With high probability, there will be $\Theta(n^{1-\beta})$ nodes accessing the same cache. The bottleneck limiting the flows from this cache to the destination nodes is the edge with capacity one connecting the cache to the tree. Hence, with this strategy, we can achieve a per-node rate of $\Theta(n^{\beta-1})$ over the graph G with high probability. By Theorem 1, this implies that a per-node rate of

$$\rho(\lambda) \geq n^{\beta-1-o(1)}$$

is achievable with probability $1-o(1)$ as $n \to \infty$ in the wireless network. A short calculation reveals that this is an order-optimal routing strategy over G, which, by Theorem 2, shows that

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

for $\alpha > 6$. Hence, for $\alpha > 6$

$$\rho(\lambda) = n^{\beta-1\pm o(1)}.$$

Moreover, it can be shown that this is the correct scaling of $\rho(\lambda)$ also for $\alpha \in (2,6]$.

This example illustrates that in situations in which the traffic demand and location of caches are regular enough, the strategy of selecting the nearest cache (as analyzed also in Example 3, and which is shown there to be arbitrarily bad in general) can actually be close to optimal. $\diamond$

IV. Proofs

In Section IV-A, we provide the proof of the inner bound in Theorem 1. The proof relies on the communication scheme presented earlier in Section III-E. The outer bound in Theorem 2 is proved in Section IV-B. It consists of two key steps, summarized by Lemmas 4 and 5 below. The first step is information-theoretic, outer bounding the caching capacity region in terms of cuts in the wireless network and then relating these cuts to cuts in the graph G. The details of this first step are provided in Section IV-C. The second step relates these cuts in the graph G to supportable flows over G. The details of this second step are provided in Section IV-D.

A. Proof of Theorem 1 (Inner Bound)

We wish to show that $\rho(\lambda) \geq b_1(n)\phi(\lambda)$ for some $b_1(n) \geq n^{-o(1)}$ uniform in λ. Equivalently, we will argue

that $\lambda \in \Lambda_G(n)$ implies $b_1(n)\lambda \in \Lambda(n)$. Assume $\lambda \in \Lambda_G(n)$; then $\phi(\lambda) \geq 1$. Let

$$f \stackrel{\triangle}{=} (f_{p,U})$$

be the corresponding solution of the linear program (2). By definition of (2), the load induced by f on each edge of G is no more than its capacity.

We now use this solution f to construct a *unicast* traffic matrix solving the caching problem. Formally, a *unicast traffic matrix* is an element $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$ associating with each source–destination pair $(u, w) \in V(n) \times V(n)$ the rate $\lambda^{\mathsf{UC}}_{u,w}$ at which destination node w requests data from source node u. The *unicast capacity region* $\Lambda(n) \subset \mathbb{R}_+^{n \times n}$ is the closure of the collection of all achievable unicast traffic matrices in the wireless network. In analogy to caching traffic, every unicast traffic matrix λ^{UC} for the wireless network induces a unicast traffic matrix between the leaf nodes of the graph G, and we can define $\Lambda^{\mathsf{UC}}_G(n)$ as the collection of unicast traffic matrices that can be routed (i.e., are supportable) over G.

Consider again the flows f as defined above. Construct the unicast traffic matrix $\lambda^{\mathsf{UC}} = \lambda^{\mathsf{UC}}(f)$ as

$$\lambda^{\mathsf{UC}}_{u,w} \stackrel{\triangle}{=} \sum_{\substack{U \subset V(n): \\ u \in U}} f_{p_{u,w},U}$$

where $p_{u,w}$ is the unique path in the tree graph G between u and w. In words, $\lambda^{\mathsf{UC}}_{u,w}$ is the sum of the flows $f_{p_{u,w},U}$ for the caching problem from u to w. The load induced by this unicast traffic $\lambda^{\mathsf{UC}}(f)$ on the edges of G is the same as that due to f. In particular, the total demand of $\lambda^{\mathsf{UC}}(f)$ across each edge is at most its capacity. Since G is a tree, this implies that $\lambda^{\mathsf{UC}}(f)$ is supportable over G, i.e., $\lambda^{\mathsf{UC}}(f) \in \Lambda^{\mathsf{UC}}_G(n)$.

We have thus transformed the problem of routing *caching* traffic over G into one of routing *unicast* traffic over G. The following result, established in [22], links the set of supportable unicast traffic matrices $\Lambda^{\mathsf{UC}}_G(n)$ over G to the unicast capacity region $\Lambda^{\mathsf{UC}}(n)$ of the wireless network.

Proposition 3: Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1'(n) \geq n^{-o(1)}$ such that

$$b_1'(n)\Lambda^{\mathsf{UC}}_G(n) \subset \Lambda^{\mathsf{UC}}(n)$$

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

Proof: See [22, Lemma 10]. ∎

Proposition 3 is established by means of an explicit communication architecture, consisting of the three layers (data layer, cooperation layer, physical layer) as described in detail in Section III-E.

Proposition 3 implies that $b_1'(n)\lambda^{\mathsf{UC}}(f) \in \Lambda^{\mathsf{UC}}(n)$. Given that the unicast traffic matrix $\lambda^{\mathsf{UC}}(f)$ was created through decomposing the caching traffic matrix λ, it follows that $b_1'(n)\lambda$ can be supported using these unicast transmissions over the wireless network. That is, $b_1(n)\lambda \in \Lambda(n)$ for

$$b_1(n) \stackrel{\triangle}{=} b_1'(n) \geq n^{-o(1)}.$$

This shows that

$$b_1(n)\Lambda_G(n) \subset \Lambda(n)$$

completing the proof of Theorem 1. ∎

B. Proof of Theorem 2 (Outer Bound)

We aim to show that

$$\rho(\lambda) \leq b_2(n)\phi(\lambda)$$

for some $b_2(n) \leq n^{o(1)}$ uniform in λ. The proof proceeds in two steps. First, we relate achievable traffic in the wireless network (characterized by $\rho(\lambda)$) to cuts in the graph G (characterized by $\hat{\rho}(\lambda)$ defined below). Second, we relate these cuts in G to supportable flows over G (characterized by $\phi(\lambda)$).

Define

$$\hat{\Lambda}(n) \stackrel{\triangle}{=} \Bigg\{ \lambda \in \mathbb{R}_+^{2^n \times n} : \sum_{U \subset S \cap V} \sum_{w \in V \setminus S} \lambda_{U,w} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v} \ \forall S \subset V_G \Bigg\}$$

with $V = V(n)$ and $V_G = V_G(n)$. Furthermore, let, for any caching traffic matrix $\lambda \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}(\lambda) \stackrel{\triangle}{=} \max\big\{\rho \geq 0 : \rho\lambda \in \hat{\Lambda}(n)\big\}. \tag{3}$$

The set $\hat{\Lambda}(n)$ corresponds to the restrictions on the set of supportable caching traffic matrices on the graph G by all possible cuts S in $V_G(n)$. Consider one such cut $S \subset V_G(n)$. For any caching traffic matrix λ that can be routed over G, the total flow

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}$$

across this cut can not be larger than the capacity of the cut

$$\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}.$$

The region $\hat{\Lambda}(n)$ is the set of caching traffic matrices satisfying all these constraints. The scalar $\hat{\rho}(\lambda)$ yields an equivalent description of $\hat{\Lambda}(n)$. Note that we can rewrite the definition of $\hat{\rho}(\lambda)$ as

$$\hat{\rho}(\lambda) = \min_{S \subset V_G(n)} \frac{\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}}{\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}}. \tag{4}$$

Recall that $\Lambda_G(n)$ is the set of supportable caching traffic matrices on G, and that $\phi(\lambda)$ is its equivalent description. From the discussion in the last paragraph, it is clear that $\Lambda_G(n) \subset \hat{\Lambda}(n)$, or, equivalently, that $\phi(\lambda) \leq \hat{\rho}(\lambda)$. The next lemma shows that $\hat{\rho}(\lambda)$ is also an approximate upper bound on the equivalent description $\rho(\lambda)$ of the caching capacity region $\Lambda(n)$ of the wireless network.

Lemma 4: Under either fast or slow fading, for any $\alpha > 6$, there exists $b_3(n) \leq n^{o(1)}$ such that

$$\rho(\lambda) \leq b_3(n)\hat{\rho}(\lambda)$$

for all caching traffic matrices $\lambda \in \mathbb{R}_+^{2^n \times n}$ with probability $1 - o(1)$ as $n \to \infty$.

The proof of Lemma 4 is presented in Section IV-C.

Lemma 4 shows that, for $\alpha > 6$, $\Lambda(n) \subset b_3(n)\hat{\Lambda}(n)$. This implication is much less obvious than the statement $\Lambda_G(n) \subset \hat{\Lambda}(n)$. The proof of Lemma 4 first uses the information-theoretic cut-set bound to upper bound achievable rates for caching traffic by cuts in the wireless network and then relates these cuts in the wireless network to cuts in the graph G. We point out that it is this step that limits the applicability of the outer bound in Theorem 2 to large path-loss exponents $\alpha > 6$. The reason for this is that evaluation of the cut-set bound for the wireless network for small path-loss exponents is quite difficult. While it is known how to evaluate "rectangular" cuts for small α [19], these techniques do not extend to the arbitrary cuts that are required for the analysis of caching traffic.

Lemma 4 allows us to upper bound the equivalent description $\rho(\lambda)$ of the caching capacity region $\Lambda(n)$ by the equivalent description $\hat{\rho}(\lambda)$ of the set $\hat{\Lambda}(n)$ of caching traffic matrices satisfying all cut constraints in the graph G. We now show that $\hat{\rho}(\lambda)$ can be upper bounded by the equivalent description $\phi(\lambda)$ of the set $\Lambda_G(n)$ of supportable caching traffic matrices on G.

Lemma 5: For any $\alpha > 2$, there exists $b_4(n) \geq n^{-o(1)}$ such that

$$b_4(n)\hat{\rho}(\lambda) \leq \phi(\lambda)$$

for all caching traffic matrices $\lambda \in \mathbb{R}_+^{2^n \times n}$.

The proof of Lemma 5 is presented in Section IV-D.

Lemma 5 shows that, for any $\alpha > 2$, $b_4(n)\hat{\Lambda}(n) \subset \Lambda_G(n)$. From the above discussion, we already know that $\Lambda_G(n) \subset \hat{\Lambda}(n)$. Hence, we deduce from Lemma 5 that $\Lambda_G(n) \approx \hat{\Lambda}(n)$. This can be understood as an approximate max-flow min-cut result for caching traffic on undirected capacitated graphs. Lemma 5 is, in fact, valid for any tree graph G (with mild assumptions on the edge capacities, see the proof for the details) and might be of independent interest.

Combining Lemmas 4 and 5 shows that, for any $\alpha > 6$

$$\begin{aligned}\rho(\lambda) &\leq b_3(n)\hat{\rho}(\lambda)\\ &\leq \frac{b_3(n)}{b_4(n)}\phi(\lambda).\end{aligned}$$

Setting

$$b_2(n) \stackrel{\triangle}{=} b_3(n)/b_4(n) \leq n^{o(1)}$$

and noting that $b_2(n)$ is uniform in λ, concludes the proof of Theorem 2.

C. Proof of Lemma 4

We start with several auxiliary results. We first introduce some regularity conditions that are satisfied with high probability by a random node placement. Define $\mathcal{V}(n)$ to be the collection of all node placements $V(n)$ that satisfy the following conditions:

$$\begin{aligned}r_{u,v} &> n^{-1}\\ &\quad \text{for all } u, v \in V(n), u \neq v\\ \big|V_{\ell,i}(n)\big| &\leq \log(n)\\ &\quad \text{for } \ell = \tfrac{1}{2}\log(n) \text{ and all } i \in \{1, \ldots, 4^\ell\}\\ \big|V_{\ell,i}(n)\big| &\geq 1\\ &\quad \text{for } \ell = \tfrac{1}{2}\log\left(\tfrac{n}{2\log(n)}\right) \text{ and all } i \in \{1, \ldots, 4^\ell\}\\ \big|V_{\ell,i}(n)\big| &\in [4^{-\ell-1}n, 4^{-\ell+1}n]\\ &\quad \text{for all } \ell \in \big\{1, \ldots, \tfrac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big)\big\},\\ &\quad i \in \{1, \ldots, 4^\ell\}.\end{aligned}$$

The first condition is that the minimum distance between node pairs is not too small. The second condition is that all squares of area 1 contain at most $\log(n)$ nodes. The third condition is that all squares of area $2\log(n)$ contain at least one node. The fourth condition is that all squares up to level $\frac{1}{2}\log(n)\big(1 - \log^{-5/6}(n)\big)$ contain a number of nodes proportional to their area.

The next lemma, quoted from [22], states that a random node placement satisfies these conditions with high probability.

Lemma 6:

$$\mathbb{P}\big(V(n) \in \mathcal{V}(n)\big) \geq 1 - o(1)$$

as $n \to \infty$.

Proof: See [22, Lemma 5]. ■

We continue with results upper bounding the MIMO capacity between subsets of nodes in $V(n)$. Formally, for disjoint subsets $S_1, S_2 \subset V(n)$, denote by $C(S_1, S_2)$ the MIMO capacity between the nodes in S_1 and S_2. Let

$$\boldsymbol{H}_{S_1,S_2} \stackrel{\triangle}{=} (h_{u,v})_{u \in S_1, v \in S_2}$$

be the matrix of channel gains between the nodes in S_1 and S_2. Under fast fading

$$C(S_1, S_2) \stackrel{\triangle}{=} \max \mathbb{E}\Big(\log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^\dagger \boldsymbol{Q}(\boldsymbol{H})\boldsymbol{H}_{S_1,S_2}\right)\Big)$$

where the maximization is over all positive semidefinite matrices $\boldsymbol{Q}(\boldsymbol{H})$ such that $\mathbb{E}(q_{u,u}) \leq 1$ for all $u \in S_1$. Under slow fading

$$C(S_1, S_2) \stackrel{\triangle}{=} \max \log\det\left(\boldsymbol{I} + \boldsymbol{H}_{S_1,S_2}^\dagger \boldsymbol{Q}\boldsymbol{H}_{S_1,S_2}\right)$$

where the maximization is over all positive semidefinite matrices $\boldsymbol{Q}$ such that $q_{u,u} \leq 1$ for all $u \in S_1$. See, e.g., [26]. To simplify notation, define furthermore

$$r_{S,v} \stackrel{\triangle}{=} \min_{u \in S} r_{u,v}$$

for $S \subset V(n)$. The next lemma provides an upper bound on the MIMO capacity $C(S, S^c)$ between the nodes in S and S^c

in terms of the number of nodes close to the boundary between them.

Lemma 7: Under either fast or slow fading, for every $\alpha > 6$, there exists a constant K_1 such that for large enough n and all $V(n) \in \mathcal{V}(n)$ and $S \subset V(n)$

$$C(S, S^c) \leq K_1 \log^4(n) \big|\{v \in S^c : r_{S,v} < \log(n) + 1\}\big|.$$

Proof: Set $S_1 \stackrel{\triangle}{=} S$ and $S_2 \stackrel{\triangle}{=} S^c$, and denote by S_2^k the nodes in S_2 that are at distance between k and $k+1$ from S_1, i.e.,

$$S_2^k \stackrel{\triangle}{=} \Big\{v \in S_2 : r_{S_1,v} \in [k, k+1)\Big\}.$$

Note that

$$S_2 = \bigcup_{k=0}^{\infty} S_2^k$$

and

$$\big|\{v \in S_2 : r_{S_1,v} < \log(n) + 1\}\big| = \sum_{k=0}^{\log(n)} |S_2^k|.$$

Applying the generalized Hadamard inequality, we obtain that under either fast or slow fading

$$C(S_1, S_2) \leq C\Big(S_1, \cup_{k=0}^{\log(n)} S_2^k\Big) + C\Big(S_1, \cup_{k>\log(n)} S_2^k\Big). \quad (5)$$

For the first term in (5), using Hadamard's inequality once more yields

$$\begin{aligned} C\Big(S_1, \cup_{k=0}^{\log(n)} S_2^k\Big) &\leq \sum_{k=0}^{\log(n)} \sum_{v \in S_2^k} C(S_1, \{v\}) \\ &\leq \sum_{k=0}^{\log(n)} \sum_{v \in S_2^k} C(\{v\}^c, \{v\}). \end{aligned}$$

By [22, Lemma 6]

$$C(\{v\}^c, \{v\}) \leq K \log(n)$$

for some constant $K < \infty$ depending only on α, and thus

$$C\Big(S_1, \cup_{k=0}^{\log(n)} S_2^k\Big) \leq K \log(n) \sum_{k=0}^{\log(n)} |S_2^k|. \quad (6)$$

For the second term in (5), we have the following upper bound from slightly adapting [13, Theorem 2.1]: Under either fast or slow fading

$$C\Big(S_1, \cup_{k>\log(n)} S_2^k\Big) \leq \sum_{k>\log(n)} \sum_{v \in S_2^k} \Big(\sum_{u \in S_1} r_{u,v}^{-\alpha/2}\Big)^2.$$

By definition of S_2^k, for $v \in S_2^k$, the (open) disk of radius k around v does not contain any node in S_1. Moreover, since $V \in \mathcal{V}$, there are at most $\log(n)$ nodes inside every subsquare of A of side length 1.[5] Thus, given that $\alpha > 6$, we have for any $v \in S_2^k$

$$\begin{aligned} \sum_{u \in S_1} r_{u,v}^{-\alpha/2} &\leq \log(n) \sum_{\tilde{k}=k}^{\infty} 10\pi(\tilde{k}+3)\tilde{k}^{-\alpha/2} \\ &\leq \tilde{K} \log(n) k^{2-\alpha/2} \end{aligned}$$

for some constant $\tilde{K} < \infty$ depending only on α. Therefore

$$C\Big(S_1, \cup_{k>\log(n)} S_2^k\Big) \leq \sum_{k>\log(n)} |S_2^k| \tilde{K}^2 \log^2(n) k^{4-\alpha}. \quad (7)$$

Consider now some $v \in S_2^k$ with $k > \log(n)$, and let u be the closest node in S_1 to v. Since $v \in S_2^k$, we must have

$$r_{u,v} \in [k, k+1).$$

Consider the (open) disk of radius $r_{u,v}$ around v and the disk of radius $\log(n)$ around u. Since u is the closest node to v in S_1, all nodes in the disk around v are in S_2. Moreover, the intersection of the two disks has an area of at least $\frac{\pi}{4}\log^2(n)$. Since $V \in \mathcal{V}$, this implies that, for n large enough, this intersection must contain at least one point, say $\tilde{v}$, and by construction

$$\tilde{v} \in \bigcup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}.$$

This shows that for every node v in S_2^k, there exists a node $\tilde{v}$ in $\cup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}$ such that

$$r_{v,\tilde{v}} \in [k - \log(n), k+1).$$

Now, since $V \in \mathcal{V}$, for every node $\tilde{v}$, there are at most

$$2\pi(k+3)(\log(n)+5)\log(n) \leq K' k \log^2(n)$$

nodes at distance $[k - \log(n), k+1)$ for some constant $K' < \infty$. Hence, the number of nodes in S_2^k is at most

$$|S_2^k| \leq K' k \log^2(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|. \quad (8)$$

Combining (8) with (7) yields

$$\begin{aligned} &C\Big(S_1, \cup_{k>\log(n)} S_2^k\Big) \\ &\quad \leq \tilde{K}^2 \log^2(n) \sum_{k>\log(n)} |S_2^k| k^{4-\alpha} \\ &\quad \leq K' \tilde{K}^2 \log^4(n) \Big(\textstyle\sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|\Big) \displaystyle\sum_{k>\log(n)} k^{5-\alpha} \\ &\quad = K'' \log^4(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}| \end{aligned} \quad (9)$$

[5] To simplify notation, we suppress dependence of $V(n), \mathcal{V}(n), \Lambda(n), \ldots$ within proofs whenever this dependence is clear from the context.

for some constant $K'' < \infty$ depending only on α, and where we have used that $\alpha > 6$. Finally, substituting (6) and (9) into (5) shows that

$$\begin{aligned} C(S_1, S_2) &\le K_1 \log^4(n) \sum_{k=0}^{\log(n)} |S_2^k| \\ &= K_1 \log^4(n) \big|\{v \in S^c : r_{S,v} < \log(n) + 1\}\big| \end{aligned}$$

with

$$K_1 \triangleq K + K''.$$

■

The next lemma shows that, for large path-loss exponents ($\alpha > 6$), every cut is approximately achievable, i.e., for every cut there exists an achievable unicast traffic matrix that has a sum rate across the cut that is not much smaller than the cut capacity.

Lemma 8: Under fast fading, for every $\alpha > 6$, there exists $b_5(n) \le n^{o(1)}$ and $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$ such that for any n, $V(n) \in \mathcal{V}(n)$, and $S \subset V(n)$

$$C(S, S^c) \le b_5(n) \sum_{u \in S} \sum_{w \notin S} \lambda_{u,w}^{\mathsf{UC}}. \tag{10}$$

Moreover, there exists a collection of channel gains $\mathcal{H}(n)$ such that

$$\mathbb{P}\big((h_{u,v}) \in \mathcal{H}(n)\big) \ge 1 - o(1)$$

as $n \to \infty$, and such that, for $(h_{u,v}) \in \mathcal{H}(n)$, (10) holds for slow fading as well.

Proof: By Lemma 7, for $V \in \mathcal{V}$

$$C(S, S^c) \le K_1 \log^4(n) \big|\{v \in S^c : r_{S,v} < \log(n) + 1\}\big|. \tag{11}$$

Construct a unicast traffic matrix $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$ as

$$\lambda_{u,w}^{\mathsf{UC}} \triangleq \begin{cases} \kappa(n), & \text{if } r_{u,w} < \log(n) + 1 \\ 0, & \text{otherwise} \end{cases}$$

for some function $\kappa(n)$. We now argue that for $\kappa(n) = \Theta(\log^{-3}(n))$ there exists $\tilde{b}(n) \ge n^{-o(1)}$ such that $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$. This follows from [22, Th. 1] (see also [22, Sec. IX.C]), once we show that for every $\ell \in \{1, \dots, L(n)\}$ and $i \in \{1, \dots, 4^\ell\}$ we have

$$\sum_{u \in V_{\ell,i}} \sum_{w \notin V_{\ell,i}} \lambda_{u,w}^{\mathsf{UC}} \le (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2} \tag{12a}$$

$$\sum_{u \notin V_{\ell,i}} \sum_{w \in V_{\ell,i}} \lambda_{u,w}^{\mathsf{UC}} \le (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2} \tag{12b}$$

and, for all $w \in V$

$$\sum_{u \ne w} \lambda_{u,w}^{\mathsf{UC}} \le 1$$

$$\sum_{u \ne w} \lambda_{w,u}^{\mathsf{UC}} \le 1.$$

Since we assume that $V \in \mathcal{V}$, we have for all $w \in V$

$$\sum_{u \ne w} \lambda_{u,w}^{\mathsf{UC}} \le K \log^3(n) \kappa(n)$$

$$\sum_{u \ne w} \lambda_{w,u}^{\mathsf{UC}} \le K \log^3(n) \kappa(n)$$

for some constant $K < \infty$. By the locality of the unicast traffic matrix λ^{UC}, it can be verified that this is sufficient for (12) to hold with

$$\kappa(n) \triangleq \frac{1}{K} \log^{-3}(n).$$

Hence [22, Th. 1] applies, showing that $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ for fast fading, and the same holds for slow fading for $\mathcal{H}$ with

$$\mathbb{P}\big((h_{u,v}) \in \mathcal{H}\big) \ge 1 - o(1)$$

as $n \to \infty$.

Now, by construction of the unicast traffic matrix λ^{UC}

$$\begin{aligned} &\sum_{u \in S} \sum_{w \notin S} \lambda_{u,w}^{\mathsf{UC}} \\ &\quad = \big|\{(u,w) \in S \times S^c : r_{u,w} < \log(n) + 1\}\big| \kappa(n) \\ &\quad \ge \big|\{w \in S^c : r_{S,w} < \log(n) + 1\}\big| \kappa(n). \end{aligned}$$

Combined with (11), this implies that

$$\begin{aligned} C(S, S^c) &\le K_1 \log^4(n) \big|\{w \in S^c : r_{S,w} < \log(n) + 1\}\big| \\ &\le \frac{K_1 \log^4(n)}{\kappa(n)} \sum_{u \in S} \sum_{w \notin S} \lambda_{u,w}^{\mathsf{UC}}. \end{aligned}$$

Since $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$, this proves the lemma with

$$b_5(n) \triangleq \frac{K_1 \log^4(n)}{\kappa(n)\tilde{b}(n)} \le n^{o(1)}.$$

■

We are now ready for the proof of Lemma 4.

Proof of Lemma 4: We wish to show that, for $\alpha > 6$, there exists $b_3(n) \le n^{o(1)}$ such that

$$\rho(\lambda) \le b_3(n) \hat{\rho}(\lambda)$$

with $\hat{\rho}(\lambda)$ as defined in (3). Consider the traffic matrix $\rho(\lambda) \cdot \lambda$ and a cut $S \subset V$ in the wireless network. Assume we allow the nodes on each side of the cut to cooperate without any restriction—this can only increase achievable rates. The total amount of traffic that needs to be transmitted across the cut is at least

$$\rho(\lambda) \sum_{U \subset S} \sum_{w \notin S} \lambda_{U,w}.$$

The maximum achievable sum rate (with the aforementioned node cooperation) is given by $C(S, S^c)$, the MIMO capacity between the nodes in S and in S^c. Therefore

$$\rho(\lambda) \le \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda_{U,w}}.$$

Since this is true for all cuts $S \subset V$, we may optimize over the choice of S to obtain the bound

$$\rho(\lambda) \leq \min_{S \subset V} \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda_{U,w}}. \tag{13}$$

We proceed by relating the cut S in the wireless network to a cut $\tilde{S}$ in the graph G. By Lemma 8, for $V \in \mathcal{V}$, there exists $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ such that for fast fading

$$C(S, S^c) \leq b_5(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w} \tag{14}$$

and (14) holds also for slow fading if $(h_{u,v}) \in \mathcal{H}$ with $\mathcal{H}$ defined as in Lemma 8. By [22, Th. 1] (see also the discussion in [22, Sec. IX.D]), for $\alpha > 6$ and $V \in \mathcal{V}$, there exists K such that if $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ then $K \log^{-6}(n) \lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_G$, where G is the tree graph defined in Section III-A.

Now, consider any $\tilde{S} \subset V_G$ such that $\tilde{S} \cap V = S$. Note that $\tilde{S}$ is a cut in G separating S from $V \setminus S$. Since $K \log^{-6}(n) \lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_G$, we thus have

$$K \log^{-6}(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in \tilde{S}, v \notin \tilde{S}}} c_{u,v}.$$

By minimizing over the choice of $\tilde{S}$ such that $\tilde{S} \cap V = S$, we obtain

$$K \log^{-6}(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w} \leq \min_{\tilde{S}: \tilde{S} \cap V = S} \sum_{\substack{(u,v) \in E_G: \\ u \in \tilde{S}, v \notin \tilde{S}}} c_{u,v}. \tag{15}$$

Combining (14) and (15) shows that

$$C(S, S^c) \leq \frac{b_5(n)}{K} \log^6(n) \min_{\tilde{S}: \tilde{S} \cap V = S} \sum_{\substack{(u,v) \in E_G: \\ u \in \tilde{S}, v \notin \tilde{S}}} c_{u,v}.$$

Together with (13), and using Lemma 6, this yields that with probability

$$\mathbb{P}((h_{u,v}) \in \mathcal{H}, V \in \mathcal{V}) \geq 1 - o(1)$$

as $n \to \infty$, we have for any caching traffic matrix λ

$$\begin{aligned}
\rho(\lambda) &\leq \min_{S \subset V} \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda_{U,w}} \\
&\leq b_3(n) \min_{S \subset V} \min_{\tilde{S} \in V_G: \tilde{S} \cap V = S} \frac{\sum_{\substack{(u,v) \in E_G: \\ u \in \tilde{S}, v \notin \tilde{S}}} c_{u,v}}{\sum_{U \subset \tilde{S} \cap V} \sum_{w \in V \setminus \tilde{S}} \lambda_{U,w}} \\
&= b_3(n) \min_{\tilde{S} \subset V_G} \frac{\sum_{\substack{(u,v) \in E_G: \\ u \in \tilde{S}, v \notin \tilde{S}}} c_{u,v}}{\sum_{U \subset \tilde{S} \cap V} \sum_{w \in V \setminus \tilde{S}} \lambda_{U,w}} \\
&= b_3(n) \hat{\rho}(\lambda)
\end{aligned}$$

with

$$b_3(n) \stackrel{\triangle}{=} \frac{b_5(n)}{K} \log^6(n) \leq n^{o(1)}$$

and where we have used (4) for the last equality. ■

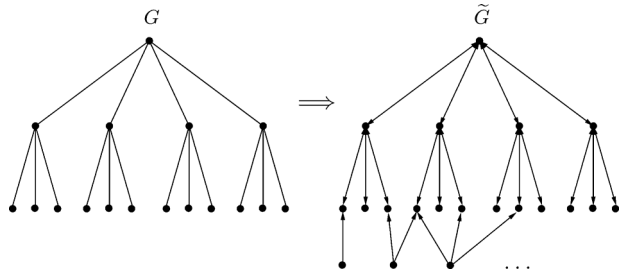

Fig. 6. Construction of the directed graph $\tilde{G}$ from the undirected graph G.

D. Proof of Lemma 5

We wish to show that there exists $b_4(n) \geq n^{-o(1)}$ such that for any λ

$$\phi(\lambda) \geq b_4(n) \hat{\rho}(\lambda) \tag{16}$$

with $\hat{\rho}(\lambda)$ as defined in (3). To this end, we need to argue that whenever a caching traffic matrix can be supported over the graph G, then there exists at least one cut in the graph that is approximately saturated. In other words, we need to argue that an approximate max-flow min-cut result holds for caching traffic over G.

The proof of the lemma proceeds as follows. We first transform the *undirected* graph G into an *directed* graph $\tilde{G}$ such that *caching* traffic can be supported over G if and only if a corresponding *unicast* traffic can be supported over $\tilde{G}$. We then argue that for unicast traffic over $\tilde{G}$ an approximate max-flow min-cut result holds. Finally, we map this result for unicast traffic on $\tilde{G}$ back to G to obtain the desired max-flow min-cut result for caching traffic over G.

Pick any $\lambda \in \mathbb{R}_+^{2^n \times n}$. For $\lambda = \mathbf{0}$, $\phi(\lambda)$ and $\hat{\rho}(\lambda)$ are both infinite, and the lemma trivially holds. Assume then that $\lambda \neq \mathbf{0}$. By rescaling λ if required, we can then assume without loss of generality that

$$\sum_{(U,w)} \lambda_{U,w} = 1. \tag{17}$$

Furthermore, we can assume that $\lambda_{U,w} = 0$ whenever $w \in U$, since then w already has access to the message it requests.

Recall that G is an *undirected* capacitated graph. We construct a *directed* capacitated graph $\tilde{G} = (V_{\tilde{G}}, E_{\tilde{G}})$ as illustrated in Fig. 6. Take the undirected graph G and turn it into a directed graph by splitting each edge $e \in E_G$ into two directed edges each with the same capacity as e. Add 2^n additional nodes to V_G, one for each subset $U \subset V(n)$. Connect the new node $\tilde{u}$ corresponding to $U \subset V(n)$ to each node $u \in U$ by a directed edge $(\tilde{u}, u)$ with infinite capacity $c_{\tilde{u},u} = \infty$.

We call the directed version of G that is contained in $\tilde{G}$ as a subgraph its *core*. Note that if some flows can be routed through G, then the same flows can be routed through the core of $\tilde{G}$, and if some flows can routed through the core of $\tilde{G}$, then at least half of each flow can be routed through G. Hence, for scaling purposes, the two are equivalent.

Now, assume we are given a caching traffic matrix λ for G. Construct a *unicast* traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ for $\tilde{G}$ by making for each

(U, w) pair in G (i.e., $U \subset V(n)$, $w \in V(n)$) the node $\tilde{u}$ in $\tilde{G}$ corresponding to U a source for w with rate

$$\tilde{\lambda}^{\mathsf{UC}}_{\tilde{u},w} \stackrel{\triangle}{=} \lambda_{U,w}.$$

For all other node pairs, the traffic demand is set to zero. Note that with this construction, all traffic over $\tilde{G}$ originates at a node in $V_{\tilde{G}} \setminus V_G$ of $\tilde{G}$ and is destined for a node in $V(n) \subset V_{\tilde{G}}$. Denote by $\Lambda^{\mathsf{UC}}_{\tilde{G}}(n)$ the set of such unicast traffic matrices that are supportable on $\tilde{G}$, and define

$$\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) \stackrel{\triangle}{=} \max\left\{\phi \geq 0 : \phi\tilde{\lambda}^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_{\tilde{G}}(n)\right\}$$

to be the equivalent description of $\Lambda^{\mathsf{UC}}_{\tilde{G}}(n)$. By construction of $\tilde{G}$ from G, and by the above argument relating G to the core of $\tilde{G}$, we have

$$\phi(\lambda) \geq \frac{1}{2}\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}). \tag{18}$$

We have thus related caching traffic in the undirected graph G to unicast traffic in the directed graph $\tilde{G}$.

We are then left with the problem of analyzing unicast traffic over $\tilde{G}$. Recall that we have seen earlier that trivially $\hat{\rho}(\lambda) \geq \phi(\lambda)$ (since the total flow over each cut can be at most equal to the cut capacity). By (18), this implies that $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) \leq 2\hat{\rho}(\lambda)$. The goal here is to establish that $\hat{\rho}(\lambda)$ is also an approximate lower bound to $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$. This is nontrivial because it requires showing that the polytope $\hat{\Lambda}(n)$ with fewer constraints is closely approximated by the polytope $\Lambda^{\mathsf{UC}}_{\tilde{G}}(n)$ with more constraints. Specifically, we are looking for this approximation to be of the form

$$b(n)\hat{\Lambda}(n) \subset \Lambda^{\mathsf{UC}}_{\tilde{G}}(n) \subset 2\hat{\Lambda}(n)$$

where $b(n) \geq n^{-o(1)}$.

In the recent literature on multicommodity flows, starting with works by Leighton and Rao [27], and by Linial *et al.* [28], such approximate max-flow min-cut results for unicast traffic for undirected graphs have been studied. However, in our context, two difficulties arise. First, $\tilde{G}$ is a directed graph. While for undirected graphs with m nodes $O(\log(m))$ approximation results for the unicast capacity region of such graphs in terms of cut-set bounds are known [28], the best known approximation result for general directed graphs is $O(m^{11/23})$ up to polylog factors in m [29]. Second, the graph $\tilde{G}$ is exponentially big in n. More precisely, $|V_{\tilde{G}}| \geq 2^n$. Hence, even a logarithmic (in the size m of the graph) approximation result will only yield a polynomial approximation in n. We are interested here in an approximation ratio that scales like $n^{o(1)}$, i.e., strictly sublogarithmic in the size of $|V_{\tilde{G}}|$. Nonetheless, as we shall see, the special structure of $\tilde{G}$ can be used to obtain an $O(\log(n)) \leq n^{o(1)}$ approximation for $\Lambda^{\mathsf{UC}}_{\tilde{G}}(n)$ in terms of $\hat{\Lambda}(n)$.

We use an idea from [30], namely that the unicast traffic problem can be reduced to a maximum sum-rate problem. More precisely, for a subset $\tilde{F} \subset V_{\tilde{G}} \times V_{\tilde{G}}$ of (u, w) pairs in $\tilde{G}$, define the *maximum sum rate* as

$$\tilde{\sigma}_{\tilde{F}} \stackrel{\triangle}{=} \max\left\{\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}} : \tilde{\lambda}^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}_{\tilde{G}}(n)\right\}$$

where here and in the following:

$$\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}} \stackrel{\triangle}{=} \sum_{(u,w)\in\tilde{F}} \tilde{\lambda}^{\mathsf{UC}}_{u,w}.$$

The quantity $\tilde{\sigma}_{\tilde{F}}$ is the largest sum rate that can be supported between the source–destination pairs in $\tilde{F}$ over the graph $\tilde{G}$.

We now argue that for every unicast traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ there exists $\tilde{F}$ such that the ratio $\tilde{\sigma}_{\tilde{F}}/\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$ is not too much bigger than $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$.

Lemma 9: Given $\tilde{\lambda}^{\mathsf{UC}}$ on $\tilde{G}$ as described above, there exists a set $\tilde{F}$ of (u, w) pairs with $u \in V_{\tilde{G}} \setminus V_G$ and $w \in V(n) \subset V_{\tilde{G}}$ so that

$$\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) \geq \frac{1}{2(1+\ln(2n^4))}\frac{\tilde{\sigma}_{\tilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}}. \tag{19}$$

Recall that the unicast traffic matrix $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}$ is the largest scalar multiple of $\tilde{\lambda}^{\mathsf{UC}}$ that is supportable over $\tilde{G}$ by definition of $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$. Hence, Lemma 9 shows that for a point $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}$ on the boundary of the region $\Lambda^{\mathsf{UC}}_{\tilde{G}}(n)$, there exists a set of source–destination pairs $\tilde{F}$ such that the total demand $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$ between the pairs in $\tilde{F}$ is almost as large as the maximum sum rate that is supportable between $\tilde{F}$. Thus, for $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}$, the pairs in $\tilde{F}$ can be understood as the approximate bottleneck, limiting further scaling of $\tilde{\lambda}^{\mathsf{UC}}$ beyond the multiple $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$. The next lemma links the ratio $\tilde{\sigma}_{\tilde{F}}/\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$ appearing in the right-hand side of (19) to the equivalent description $\hat{\rho}(\lambda)$ of the region $\hat{\Lambda}(n)$.

Lemma 10: For any set $\tilde{F}$ of (u, w) pairs with $u \in V_{\tilde{G}} \setminus V_G$ and $w \in V(n) \subset V_{\tilde{G}}$

$$\frac{\tilde{\sigma}_{\tilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}} \geq \frac{1}{4}\hat{\rho}(\lambda).$$

Combining Lemmas 9 and 10 with (18) shows that

$$\begin{aligned}\phi(\lambda) &\geq \frac{1}{2}\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) \\ &\geq \frac{1}{4(1+\ln(2n^4))}\frac{\tilde{\sigma}_{\tilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}} \\ &\geq \frac{1}{16(1+\ln(2n^4))}\hat{\rho}(\lambda).\end{aligned}$$

This establishes Lemma 5 with

$$b_4(n) \stackrel{\triangle}{=} \frac{1}{16(1+\ln(2n^4))} \geq n^{-o(1)}.$$

It remains to prove Lemmas 9 and 10.

Proof of Lemma 9: Given a unicast traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ on $\tilde{G}$ as described above, we want to find a set of node pairs $\tilde{F}$ such that $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$ is not too much smaller than the ratio $\tilde{\sigma}_{\tilde{F}}/\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$.

First, note that $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$ is the solution to the following linear program:

$$\begin{aligned}\max \quad & \phi \\ \text{s.t.} \quad & \textstyle\sum_{p\in\tilde{P}_{u,w}} f_p \geq \phi\tilde{\lambda}^{\mathsf{UC}}_{u,w} && \forall u, w \in V_{\tilde{G}} \\ & \textstyle\sum_{p\in\tilde{P}:e\in p} f_p \leq c_e && \forall e \in E_{\tilde{G}} \\ & f_p \geq 0 && \forall p \in \tilde{P}\end{aligned} \tag{20}$$

where $\tilde{P}_{u,w}$ is the collection of all paths in $\tilde{G}$ from node u to node w, and

$$\tilde{P} \stackrel{\triangle}{=} \bigcup_{(u,w)\in V_{\tilde{G}}\times V_{\tilde{G}}} \tilde{P}_{u,w}.$$

The corresponding dual linear program is

$$\begin{array}{lrll} \min & \sum_{e\in E_{\tilde{G}}} c_e m_e & & \\ \text{s.t.} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall u,w \in V_{\tilde{G}}, p \in \tilde{P}_{u,w} \\ & \sum_{u,w\in V_{\tilde{G}}} d_{u,w}\tilde{\lambda}^{\mathsf{UC}}_{u,w} & \geq 1 & \\ & m_e & \geq 0 & \forall e \in E_{\tilde{G}} \\ & d_{u,w} & \geq 0 & \forall u,w \in V_{\tilde{G}}. \end{array} \tag{21}$$

Since the all-zero solution is feasible for the primal program (20), strong duality holds, i.e., the maximum in the primal (20) is equal to the minimum in the dual (21). Moreover, by weak duality, any feasible solution to the dual problem (21) yields an upper bound to the maximum in the primal (20).

Second, $\tilde{\sigma}_{\tilde{F}}$ is the solution to the linear program

$$\begin{array}{lll} \max \sum_{(u,w)\in\tilde{F}} \sum_{p\in\tilde{P}_{u,w}} f_p & & \\ \text{s.t.} \quad \sum_{p\in\tilde{P}:e\in p} f_p \leq c_e & \forall e \in E_{\tilde{G}} \\ f_p \geq 0 & \forall p \in \tilde{P} \end{array} \tag{22}$$

and its dual is

$$\begin{array}{lrll} \min & \sum_{e\in E_{\tilde{G}}} c_e m_e & & \\ \text{s.t.} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall u,w \in V_{\tilde{G}}, p \in \tilde{P}_{u,w} \\ & d_{u,w} & \geq 1 & \forall (u,w) \in \tilde{F} \\ & m_e & \geq 0 & \forall e \in E_{\tilde{G}} \\ & d_{u,w} & \geq 0 & \forall u,w \in V_{\tilde{G}}. \end{array} \tag{23}$$

Again strong and weak duality hold.

Let $(m^\star_e)_{e\in E_{\tilde{G}}}$, $(d^\star_{u,w})_{u,w\in V_{\tilde{G}}}$ be a minimizer for the dual (21) of the unicast traffic problem. By strong duality, the minimum of the dual (21) is equal to the maximum of the corresponding primal (20). We now show how $(m^\star_e)$, $(d^\star_{u,w})$ can be used to construct a feasible solution to the dual (23) of the maximum sum-rate problem for a specific choice of subset $\tilde{F}$. By weak duality, this feasible solution for the dual (23) yields an upper bound on the maximum in the corresponding primal (22). This will allow us to lower bound $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$ in terms of the ratio $\tilde{\sigma}_{\tilde{F}}/\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$ as required.

Note first that we can assume without loss of optimality that

$$d^\star_{u,w} = \begin{cases} 0, & \text{if } \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 0 \\ \min_{p\in\tilde{P}_{u,w}} \sum_{e\in p} m^\star_e, & \text{otherwise.} \end{cases} \tag{24}$$

Now, since $c_e = \infty$ whenever $e \in E_{\tilde{G}} \setminus E_G$, we have $m^\star_e = 0$ for those edges. Since, in addition, $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$ only if $u \in V_{\tilde{G}} \setminus V_G$ and if w is a leaf node of G, this implies that $(d^\star_{u,w})_{u,w\in V_{\tilde{G}}}$ can take at most n^2 different nonzero values, since there are at most that many distinct paths between leaf nodes in the tree graph G. Order these values in decreasing order

$$d^\star_1 > d^\star_2 > \cdots > d^\star_K > d^\star_{K+1} = 0$$

with $K \leq n^2$, and define for $1 \leq k \leq K$

$$\tilde{\lambda}^{\mathsf{UC}}_k \stackrel{\triangle}{=} \sum_{u,w\in V_{\tilde{G}}: d^\star_{u,w} = d^\star_k} \tilde{\lambda}^{\mathsf{UC}}_{u,w}. \tag{25}$$

We now argue that $d^\star_k \leq n^2$ for all $k \in \{1,\ldots,K\}$. In fact, assume $d^\star_1 > n^2$, then by (24) there exists at least one edge $\tilde{e}$ such that $m^\star_{\tilde{e}} > n$, because in any path $P_{u,w}$, there are at most n edges with non-zero $m^\star_{\tilde{e}}$ value. Hence

$$\sum_{e\in E_{\tilde{G}}} c_e m^\star_e \geq c_{\tilde{e}} m^\star_{\tilde{e}} > n$$

since $c_e \geq 1$ for all $e \in E_{\tilde{G}}$. Due to strong duality, this implies that the solution of the linear program (20), i.e., the value of $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$, is strictly larger than n. But that is not possible. Indeed, due to the normalization assumption (17), we have $\sum_{u,w\in V_{\tilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 1$. By construction, all destination nodes w in $\tilde{V}_G$ are in $V \subset V_{\tilde{G}}$, and hence, there are at most n nodes w with nonzero $\tilde{\lambda}^{\mathsf{UC}}_{u,w}$. Together, this implies that for at least one node w the total traffic into w satisfies

$$\sum_{u\in V_{\tilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \geq \frac{1}{n}.$$

By definition, $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}$ must be supportable in $\tilde{G}$. Since $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) > 0$, and since, by assumption, $\tilde{\lambda}^{\mathsf{UC}}_{U,w} = 0$ whenever $w \in U$, this will induce a load strictly greater than one on the finite capacity edge incident on w. As $w \in V$, this edge has unit capacity, which contradicts that $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})\tilde{\lambda}^{\mathsf{UC}}$ is supportable. Therefore, $\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$ must be no more than n, and hence, we obtain that $d^\star_k \leq d^\star_1 \leq n^2$ for all $1 \leq k \leq K$.

We now argue that at least one of $d^\star_k$ in $1 \leq k \leq K$ is not too small. To that end, let $k_1 < k_2 < \cdots < k_I$ be such that

$$\{k_i\}_{i=1}^I = \left\{k : \tilde{\lambda}^{\mathsf{UC}}_k \geq \frac{1}{2n^4}\right\} \tag{26}$$

with $\tilde{\lambda}^{\mathsf{UC}}_k$ as defined in (25). Note that $I \geq 1$ since otherwise

$$\begin{aligned} \sum_{u,w\in V_{\tilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} &= \sum_{k=1}^K \tilde{\lambda}^{\mathsf{UC}}_k \\ &< \frac{K}{2n^4} \\ &\leq 1 \end{aligned}$$

contradicting the normalization assumption (17). Define

$$s_i \stackrel{\triangle}{=} \sum_{j=1}^i \tilde{\lambda}^{\mathsf{UC}}_{k_j}.$$

Using that $(d^\star_k)$ is feasible for the dual (21), that $d^\star_k \leq n^2$, and that $K \leq n^2$, we have

$$\begin{aligned} \sum_{i=1}^I d^\star_{k_i} \tilde{\lambda}^{\mathsf{UC}}_{k_i} &\geq 1 - \sum_{k:\tilde{\lambda}^{\mathsf{UC}}_k < 1/2n^4} d^\star_k \tilde{\lambda}^{\mathsf{UC}}_k \\ &\geq 1 - Kn^2 \frac{1}{2n^4} \\ &\geq \frac{1}{2}. \end{aligned} \tag{27}$$

We argue that this implies existence of i such that

$$d^\star_{k_i} \geq \frac{1}{2s_i(1+\ln(2n^4))}. \tag{28}$$

Indeed, assume (28) is false for all i. Then

$$\begin{aligned}
&\sum_{i=1}^{I} d^\star_{k_i} \tilde{\lambda}^{\mathsf{UC}}_{k_i} \\
&\quad < \frac{1}{2(1+\ln(2n^4))} \sum_{i=1}^{I} \frac{\tilde{\lambda}^{\mathsf{UC}}_{k_i}}{s_i} \\
&\quad \overset{(a)}{=} \frac{1}{2(1+\ln(2n^4))} \Big(1 + \textstyle\sum_{i=2}^{I} \frac{s_i - s_{i-1}}{s_i}\Big) \\
&\quad \overset{(b)}{\leq} \frac{1}{2(1+\ln(2n^4))} \Big(1 + \textstyle\sum_{i=2}^{I} \big(\ln(s_i) - \ln(s_{i-1})\big)\Big) \\
&\quad = \frac{1}{2(1+\ln(2n^4))} \Big(1 + \ln\Big(s_I / \tilde{\lambda}^{\mathsf{UC}}_{k_1}\Big)\Big) \\
&\quad \overset{(c)}{\leq} \frac{1}{2(1+\ln(2n^4))} \big(1 + \ln(2n^4)\big) \\
&\quad = \frac{1}{2}
\end{aligned}$$

where we have used that $I \geq 1$ in (a), that $1 - x \leq -\ln(x)$ for every $x \geq 0$ in (b), and that $s_I \leq 1$ by (17) and $\tilde{\lambda}^{\mathsf{UC}}_{k_1} \geq \frac{1}{2n^4}$ in (c). This contradicts (27), showing that (28) must hold for some i. Consider this value of i in the following.

Now, consider the following set $\tilde{F}$ of (u,w) pairs:

$$\tilde{F} \overset{\triangle}{=} \big\{(u,w) : d^\star_{u,w} \geq d^\star_{k_i}\big\}.$$

Note that, by (24), $\tilde{F}$ contains only pairs (u,w) such that $u \in V_{\tilde{G}} \setminus V_G$ and $w \in V \subset V_{\tilde{G}}$ (i.e., nodes in $\tilde{G}$ corresponding to leaf nodes in G). Set

$$\begin{aligned}
d_{u,w} &\overset{\triangle}{=} \frac{d^\star_{u,w}}{d^\star_{k_i}} \\
m_e &\overset{\triangle}{=} \frac{m^\star_e}{d^\star_{k_i}}
\end{aligned}$$

for all $u, w \in V_{\tilde{G}}$. Note that, for $(u,w) \in \tilde{F}$

$$d_{u,w} = \frac{d^\star_{u,w}}{d^\star_{k_i}} \geq 1$$

and that for all $u, w \in V_{\tilde{G}}$, $p \in \tilde{P}_{u,w}$

$$\begin{aligned}
\sum_{e \in p} m_e &= \frac{1}{d^\star_{k_i}} \sum_{e \in p} m^\star_e \\
&\geq \frac{1}{d^\star_{k_i}} d^\star_{u,w} \\
&= d_{u,w}
\end{aligned}$$

by feasibility of $(d^\star_{u,w})$ and $(m^\star_e)$ for the dual (21). Hence, for this $\tilde{F}$, the choice of (m_e) and $(d_{u,w})$ is feasible for the dual (23). By weak duality, any feasible solution for the dual (23) yields an upper bound for the corresponding primal (22). Therefore

$$\begin{aligned}
\tilde{\sigma}_{\tilde{F}} &\leq \sum_{e \in E_{\tilde{G}}} c_e m_e \\
&= \frac{1}{d^\star_{k_i}} \sum_{e \in E_{\tilde{G}}} c_e m^\star_e.
\end{aligned}$$

By (28)

$$d^\star_{k_i} \geq \frac{1}{2s_i(1+\ln(2n^4))}$$

and, since $d^\star_{k_j} \geq d^\star_{k_i}$ for all $j \leq i$

$$\begin{aligned}
s_i &= \sum_{j=1}^{i} \tilde{\lambda}^{\mathsf{UC}}_{k_j} \\
&= \sum_{j=1}^{i} \sum_{(u,w): d^\star_{u,w} = d^\star_{k_j}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\
&\leq \sum_{(u,w): d^\star_{u,w} \geq d^\star_{k_i}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\
&= \sum_{(u,w) \in \tilde{F}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\
&= \tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}
\end{aligned}$$

(note that the last equality is simply the definition of $\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}$). Therefore

$$\begin{aligned}
\tilde{\sigma}_{\tilde{F}} &\leq \frac{1}{d^\star_{k_i}} \sum_{e \in E_{\tilde{G}}} c_e m^\star_e \\
&\leq 2s_i(1+\ln(2n^4)) \sum_{e \in E_{\tilde{G}}} c_e m^\star_e \\
&\leq 2\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}(1+\ln(2n^4)) \sum_{e \in E_{\tilde{G}}} c_e m^\star_e.
\end{aligned}$$

Since, by assumption, $(m^\star_e)$ is optimal for the dual (21), and by strong duality, we have

$$\sum_{e \in E_{\tilde{G}}} c_e m^\star_e = \tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}})$$

and hence

$$\tilde{\phi}(\tilde{\lambda}^{\mathsf{UC}}) \geq \frac{1}{2(1+\ln(2n^4))} \frac{\tilde{\sigma}_{\tilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\tilde{F}}}.$$

■

Proof of Lemma 10: We wish to analyze maximum sum rates $\tilde{\sigma}_{\tilde{F}}$ in $\tilde{G}$ for sets $\tilde{F}$ such that for $(\tilde{u}, w) \in \tilde{F}$ we have $\tilde{u} \in V_{\tilde{G}} \setminus V_G$ and $w \in V \subset V_G \subset V_{\tilde{G}}$. Notice that, due to this form of $\tilde{F}$ and since the edges in $E_{\tilde{G}} \setminus E_G$ have infinite capacity, this analysis can be done by considering only the core of $\tilde{G}$. More precisely, for a collection of node pairs $\tilde{F}$ in $\tilde{G}$ as above, we construct a collection of node pairs F in G as follows. For each $(\tilde{u}, w) \in \tilde{F}$, note that by construction $\tilde{u}$ is connected to a subset

$U \subset V \subset V_G \subset V_{\tilde{G}}$ of nodes. For each $(\tilde{u}, w) \in \tilde{F}$, add (u, w) to F for each such $u \in U$. Denote by σ_F the maximum sum rate for F in G. Since G is the undirected version of the core of $\tilde{G}$, we have

$$\tilde{\sigma}_{\tilde{F}} \geq \sigma_F. \tag{29}$$

For a collection of node pairs F in G, we call a set of edges M a *multicut* for F if in the graph $(V_G, E_G \setminus M)$ each pair in F is disconnected. For a subset $M \subset E_G$, define

$$c_M \stackrel{\triangle}{=} \sum_{e \in M} c_e.$$

From the definition of a multicut, it follows directly that $\sigma_F \leq c_M$. More surprisingly, it is shown in [31, Th. 8] that if G is an undirected tree, then for every $F \in V_G \times V_G$ there exists a multicut M for F such that

$$\sigma_F \geq \frac{1}{2} c_M. \tag{30}$$

Next, we show how the edge cut $M \subset E_G$ can be transformed into a node cut $S \subset V_G$. Denote by $\{S_i\}$ the connected components of $(V_G, E_G \setminus M)$. We can assume without loss of generality that

$$M = \bigcup_i (S_i \times S_i^c) \cap E_G$$

since otherwise we can remove the additional edges from M to create a smaller multicut for F. We, therefore, have

$$c_M = \frac{1}{2} \sum_i c_{(S_i^c \times S_i) \cap E_G} \tag{31}$$

since every edge in M appears exactly twice in the sum on the right-hand side. Define for $S \subset V_G$

$$\lambda_S \stackrel{\triangle}{=} \sum_{U \subset S \cap V} \sum_{w \in V \setminus S} \lambda_{U,w}$$

as the total caching traffic that needs to be transmitted between $S \cap V$ and $V \setminus S$. M is a multicut for F induced by $\tilde{F}$, and hence for every $(\tilde{u}, w) \in \tilde{F}$ and the corresponding pair (U, w), M separates w from all the nodes in U. Therefore, for each such (U, w) pair, there exists a set S_i such that $w \in S_i, U \subset S_i^c$. This shows that

$$\tilde{\lambda}_{\tilde{F}}^{\mathsf{UC}} \leq \sum_i \lambda_{S_i^c}. \tag{32}$$

Equations (30)–(32) imply that there exists j such that

$$\begin{aligned} \frac{\tilde{\sigma}_{\tilde{F}}}{\tilde{\lambda}_{\tilde{F}}^{\mathsf{UC}}} &\geq \frac{1}{4} \frac{\sum_i c_{(S_i^c \times S_i) \cap E_G}}{\sum_i \lambda_{S_i^c}} \\ &\geq \frac{1}{4} \frac{c_{(S_j^c \times S_j) \cap E_G}}{\lambda_{S_j^c}} \\ &\geq \frac{1}{4} \min_{S \subset V_G} \frac{c_{(S \times S^c) \cap E_G}}{\lambda_S} \\ &= \frac{1}{4} \hat{\rho}(\lambda) \end{aligned}$$

where in the last equality we have used (4). This completes the proof of Lemma 10. ■

V. Conclusion

We have analyzed the influence of caching on the performance of wireless networks. Our approach is information-theoretic, yielding an inner bound on the caching capacity region for all values $\alpha > 2$ of path-loss exponent, and a matching (in the scaling sense) outer bound for $\alpha > 6$. Thus, in the high path-loss regime $\alpha > 6$, this provides a scaling characterization of the complete caching capacity region. Even though this region is $2^n \times n$-dimensional, i.e., exponential in the number of nodes n in the wireless network, we have presented an algorithm that checks approximate feasibility of a particular caching traffic matrix efficiently, namely in polynomial time in the description length of the caching traffic matrix. Achievability is proved using a three-layer communication architecture. The three layers deal with optimal selection of caches, choice of amount of necessary cooperation, noise, and interference, respectively. The matching (again in the scaling sense) converse proves that addressing these questions separately is without loss of order-optimality in the high path-loss regime. That is, source-channel separation is close to optimal for caching traffic in this regime.

We view this result as a step toward understanding the performance loss incurred due to source-channel separation for the transmission of arbitrarily correlated sources. Determining the performance loss for such a separation based strategy for all values of $\alpha > 2$ for caching traffic and more generally for sources with arbitrary correlation are interesting questions for future research.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments.

References

[1] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

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

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

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

[5] B.-J. Ko and D. Rubenstein, "Distributed self-stabilizing placement of replicated resources in emerging networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 3, pp. 476–487, Jun. 2005.

[6] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 1, pp. 77–89, Jan. 2005.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Oct. 1948.

[8] T. Cover, A. E. Gamal, and M. Salehi, "Multiple access channels with arbitrarily correlated sources," *IEEE Trans. Inf. Theory*, vol. 26, no. 6, pp. 648–657, Nov. 1980.

[9] T. S. Han, "Slepian-Wolf-Cover theorem for network of channels," *Inf. Control*, vol. 47, no. 1, pp. 67–83, Jan. 1980.

[10] J. Barros and S. D. Servetto, "Network information flow with correlated sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 155–170, Jan. 2006.

[11] C. Tian, J. Chen, S. Diggavi, and S. Shamai, "Optimality and approximate optimality of source-channel separation in networks," [Online]. Available: arXiv:1004.2648 (cs.IT), Apr. 2010

[12] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 748–767, May 2004.
[13] A. Jovicic, P. Viswanath, and S. R. Kulkarni, "Upper bounds to transport capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2555–2565, Nov. 2004.
[14] O. Léveque and I. E. Telatar, "Information-theoretic upper bounds on the capacity of large extended ad hoc wireless networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 858–865, Mar. 2005.
[15] F. Xue, L.-L. Xie, and P. R. Kumar, "The transport capacity of wireless networks over fading channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 834–847, Mar. 2005.
[16] 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*, vol. 52, no. 6, pp. 2313–2328, Jun. 2006.
[17] 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*, vol. 53, no. 3, pp. 1009–1018, Mar. 2007.
[18] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: An achievable rate region," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1877–1894, Aug. 2003.
[19] A. Özgür, O. Léveque, and D. N. C. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3549–3572, Oct. 2007.
[20] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3959–3982, Sep. 2009.
[21] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3413–3424, Aug. 2009.
[22] U. Niesen, P. Gupta, and D. Shah, "The balanced unicast and multicast capacity regions of large wireless networks," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2249–2271, May 2010.
[23] S.-H. Lee and S.-Y. Chung, "On the capacity scaling of wireless networks: Effect of finite wavelength," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1713–1717.
[24] A.Özgür, O. Léveque, and D. N. C. Tse, "Linear capacity scaling in wireless networks: Beyond physical limits?," in *Proc. ITA*, Jan. 2010, pp. 1–10.
[25] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proc. ACM Symp. Theory Comput.*, Apr. 1984, pp. 302–311.
[26] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[27] F. T. Leighton and S. Rao, "Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms," *J. ACM*, vol. 46, no. 6, pp. 787–832, Nov. 1999.
[28] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, Jun. 1995.
[29] A. Agarwal, N. Alon, and M. Charikar, "Improved approximation for directed cut problems," *Proc. ACM Symp. Theory Comput.*, pp. 671–680, Jun. 2007.
[30] N. Kahale, Unpublished Manuscript, On reducing the cut ratio to the multicut problem. 1993.
[31] N. Garg, V. V. Vazirani, and M. Yannakakis, *Primal-Dual Approximation Algorithms for Integral Flow and Multicut in Trees, With Applications to Matching and set Cover*, ser. ser. Lecture Notes in Computer Science. New York: Springer, 1993, pp. 64–75.

Urs Niesen (S'03–M'09) received the M.S. degree from the School of Computer and Communication Sciences at theÉcole Polytechnique Fédérale de Lausanne (EPFL) in 2005 and the Ph.D. degree from the department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT) in 2009. Currently, he is a member of technical staff in the Mathematics of Networks and Communications Research Department at Bell Labs, Alcatel-Lucent. His research interests are in the areas of communication and information theory.

Devavrat Shah (M'05) is currently a Jamieson career development associate professor with the department of electrical engineering and computer science, MIT. He is a member of the Laboratory for Information and Decision Systems (LIDS) and Operations Research Center (ORC). His research focus is on theory of large complex networks which includes network algorithms, stochastic networks, network information theory and large scale statistical inference.

Devavrat Shah received his Bachelor of Technology in Computer Science and Engineering from Indian Institute of Technology, Bombay in 1999 with the Presidents of India Gold Medal—awarded to the best graduating student across all engineering disciplines. He received his PhD in Computer Science from Stanford University in 2004. His doctoral thesis titled "Randomization and Heavy Traffic Theory: New Approaches for Switch Scheduling Algorithms" was completed under supervision of Balaji Prabhakar. His thesis was adjudged winner of George B. Dantzig best dissertation award from INFORMS in 2005. After spending a year between Stanford, Berkeley and MSRI, he started teaching at MIT in Fall 2005.

Devavrat Shah has been co-awarded best paper awards at the IEEE INFOCOM '04, ACM SIGMETRICS/Performance '06; and he has supervised best student paper awards at Neural Information Processing Systems '08, ACM SIGMETRICS/Performance '09 and Management Science and Operations Management Paper competition '10.

He was awarded the first ACM SIGMETRICS Rising Star Award 2008 for his work on network scheduling algorithms. He received the 2010 Erlang Prize from INFORMS which is given to a young researcher for outstanding contributions to applied probability. He is currently an associate editor of Operations Research.

Gregory W. Wornell (S'83–M'91–SM'00–F'04) received the B.A.Sc. degree from the University of British Columbia, Vancouver, BC, Canada, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, all in electrical engineering and computer science, in 1985, 1987, and 1991, respectively.

Since 1991, he has been on the faculty at MIT, where he is Professor of Electrical Engineering and Computer Science (EECS), leads the Signals, Information, and Algorithms Laboratory in the Research Laboratory of Electronics, chairs the graduate area of Communication, Control, and Signal Processing in EECS, and is Co-Director of the Center for Wireless Networking. He has held visiting appointments at the former AT&T Bell Laboratories, Murray Hill, NJ, the University of California, Berkeley, CA, and Hewlett-Packard Laboratories, Palo Alto, CA. His research interests and publications span the areas of signal processing, digital communication, and information theory, and include algorithms and architectures for wireless and sensor networks, broadband systems, and multimedia environments.

Dr. Wornell has been involved in the Information Theory and Signal Processing Societies of the IEEE in a variety of capacities, and maintains a number of close industrial relationships and activities. He has won a number of awards for both his research and teaching.