On Downlink Capacity of Cellular Data Networks with WLAN/WPAN Relays

Božidar Radunović, *Member, IEEE,* Alexandre Proutiere, *Member, IEEE,*

***Abstract*—We consider the downlink of a cellular network supporting data traffic in which each user is equipped with the same type of 802.11-like WLAN or WPAN interface, used to relay packets to further users. We are interested in the design guidelines for such networks and how much capacity improvements can the additional relay layer bring. A first objective is to provide a scheduling/relay strategy that maximizes the network capacity. Using theoretical analysis, numerical evaluation and simulations, we find that, when the number of active users is large, the capacity-achieving strategy divides the cell into two areas: one closer to the base-station where the relay layer is always saturated and some nodes receive traffic through both direct and relay links, and the further one where the relay is never saturated and the direct traffic is almost nonexistent. We also show that it is approximately optimal to use fixed relay link lengths, and we derive this length. We show that the obtained capacity is independent of the cell size (unlike in traditional cellular networks). Based on our findings we propose simple, decentralized routing and scheduling protocols. We show that in a fully saturated network our optimized protocol substantially improves performance over the protocols that use naive relay-only or direct-only policies.**

I. Introduction

A. Cellular Networks with Relays

Wireless cellular networks operate on expensive licensed frequencies and their bandwidth is a scarce resource limited by regulations. Recently there has been a lot of interest in increasing the capacity of cellular networks using an additional wireless physical layer that operates on an unlicensed frequency band. New generations of mobile devices are already equipped with WLAN (wireless local-area network) or WPAN (wireless personal-area network) interfaces and the question that arises is whether one can use the available relay structure to improve the service of cellular networks.

In this paper we consider such a scenario and assume that mobile nodes and the base-station (BS) are equipped with an additional relay adapter. The BS can communicate with a mobile node using a *direct* link (transmission over the cellular, high-power, expensive frequency) or relaying over one or several mobile nodes using *relay* links (over the unlicensed, low-power frequency). A direct and a relay link use different frequencies, hence they can be used simultaneously by a node.

A typical relay technology we have in mind is 802.11 WLAN. The physical layer of 802.11 allows a source and a destination to adapt their communication rate. A source selects an appropriate rate to transmit a packet depending on the link quality and the level of interference at the receiver. If the link quality degrades during the packet transmission, the packet is lost and has to be retransmitted. In order to guarantee some link quality, 802.11 MAC introduces the RTS/CTS mechanism that prevents nodes in the neighborhood to interfere with an ongoing transmission. The size of this exclusion area depends on the transmission power of RTS and CTS signaling packets. Many of the existing WLAN and WPAN technologies (e.g. 802.11, 802.15.4) are based on the design principles described above. In this work we shall consider technologies using these principles. Furthermore, we assume all nodes possess the same type of relay interface.

In this paper, we consider downlink data traffic only (uplink traffic requires a different analysis; see e.g. [1]). The downlink traffic has to be carried from the BS to the users, either using direct transmissions from the BS or relays capabilities. The key component of the system is then the resource allocation strategy, which consists of a scheduling scheme sharing the BS resources, and a routing / scheduling scheme to exploit relay capabilities. Our goal is to design optimal scheduling and relay strategies.

B. Related work

Augmenting a cellular network with relays is not a novel concept. Some of the first papers that proposed this kind of architecture are [2], [3], [4]. In [2], the authors suppose that mobile nodes cannot relay and introduce dedicated relays which use unlicensed frequencies in order to improve the capacity. In [3], the authors assume mobile nodes themselves dispose of WLAN interfaces, and provide a routing protocol that finds and maintains relay routes. In [4], [5], small networks with 1-hop relays are considered.

Scheduling algorithms for relay networks are discussed in [6], [7]. In [6], the authors discuss several simple scheduling schemes. More advanced scheduling techniques are considered in [7]. There, as opposed to the other related work, it is supposed that the BS and the relays use the same frequency band. Consequently, the BS transmits only to the nearest nodes, and the others receive relay traffic only. Relaying for uplink traffic is considered in [1] which discusses a similar routing problem.

What is common for all the proposed relay protocols is that none of them is based on the objective to maximize a certain network-wide performance criterion. Instead, they are based on a simple local heuristic that considers relaying only for those nodes whose direct communication with the BS is of very low quality. That way one node will never receive traffic from *both* relay and direct links. Typically, closer nodes will receive traffic *only* directly, and distant nodes *only* over relay links.

B. Radunović is with Microsoft Research, Cambridge. The work was done during his stay at ENS. A. Proutiere is with KTH.

A large number of papers analyze the optimal resource allocation schemes for multi-hop wireless networks (e.g. [8], [9], [10], [11]). The underlying optimization problem has exponential complexity and the results cannot be directly used for implementation purposes. A cellular network with relays can be regarded as a special case of a multi-hop wireless network. We use its specific structure to simplify some of the proposed models and algorithms for multi-hop networks.

C. Contributions

In this paper, we wish to propose resource allocation strategies that combine high efficiency and manageable complexity. Specifically we characterize resource allocation strategies that maximize the network *capacity* defined as a weighted total throughput of a cell – refer to Section II.B for a precise definition. We formulate this problem as an optimization problem.

The optimization problem is difficult to solve in general. We first focus on a one-dimensional problem. We find an upper bound on the optimal solution of the one-dimensional problem (Section III-B), with several simplifying assumptions (fixed link lengths, no fading, large number of users). We show that this upper bound corresponds to the optimal strategy when the traffic is equally spread across the cell.

We next turn to a two-dimensional case, and we consider the same simplifying assumptions as previously (fixed link lengths, no fading, large number of users). With an additional approximation on the scheduling constraints (Section III-D) we are able to find an upper bound on the optimal solution of the problem in a 2D scenario.

From the above optimization problem and the structure of the solutions we obtain two important protocol design insights

- We show that the upper bound on the capacity remains *constant*, independent of the cell size, which is in contrast with cellular networks with no relay where the capacity *decreases exponentially*.
- The optimal scheduling (*MaxRelay*, illustrated in Figure 1) divides a cell into two regions. The first region, around the BS, is such that the relay channel is fully saturated. Nodes in this region may receive traffic *both* from relays and directly from the BS, which contrasts with previously proposed relay protocols [7], [5], [3], [4]. In the other region, the relay channel is never saturated, and there is no direct traffic to users.

Inspired by the key characteristics of the optimal resource sharing strategy obtained in the simplified scenario, we design simple decentralized scheduling and routing schemes that perform very well in more realistic network conditions (channels with fading, finite number of nodes). We first verify numerically that it is indeed almost optimal in the one-dimensional problem to use links of fixed length, and we conjecture that this fixed link length corresponds to the one that maximizes transport capacity (Section III-C), both in the 1D and the 2D cases. We use these finding to build a routing algorithm.

Although our scheduling and routing schemes are in no way optimal, we show that in a uniformly loaded, saturated network they offer a significant improvement in terms of capacity as compared to the conventional direct transmissions and to a naive, relay-only case when the direct traffic is scheduled only to the nodes nearest to the BS (this mimics [7] for two frequencies). This improvement is even larger when a low-rate WPAN is used instead of WLAN as a relay network.

D. Organization of the Paper

In the following section, we precisely define the modeling assumptions and the performance objectives. In Section III, we characterize the resource allocation schemes maximizing the capacity of networks with relays under several simplifying assumptions (ideal scheduling, no fading, large number of nodes). We first analyze 1D networks with fixed relay link lengths in Section III-B, we extend the analysis to the case of variable relay link lengths in Section III-C, and in Section III-D we present a heuristic to generalize our findings to 2D networks. We propose a decentralized routing and scheduling protocol in Section IV and we evaluate its performance in Section V.

II. Models and Objectives

We consider the downlink of a single cell whose transmission resources (power and bandwidth) are shared by a fixed population of data flows. Each flow is characterized by the position of the corresponding user. Denote by $\mathcal{C}$ the set of locations in the cell, and by $\mathcal{N} \subset \mathcal{C}$ the set of locations of users with active flows. Without loss of generality, we assume no two users are at the exactly same location. We will consider both 1D linear or 2D cells.

Variable	*Explanation*
$\tau(x)$	fraction of time BS servers location x
$\tau_r(s,d)$	fraction of time the node at location s relays to the node at location d
$C_d(x)$	PHY rate at which BS serves location x
$t(x) = 1/C_d(x)$	the time needed to send a unit of data to location x
$C_r(D_r)$	PHY rate of a relay link of length D_r
$\phi_d(x) = C_d(x)\tau(x)$	average rate at which BS serves location x
$\phi_r(s,d) = C_r(d-s)\tau_r(s,d)$	average rate at which the node at location s relays to the node at location d
ρ	cell throughput
$p(x)$	fraction of cell throughput received at location x
$\rho(x) = \rho p(x)$	total traffic destined to the user at location x.

TABLE I
List of variables used in the models

A. Radio Resources

We next describe the two types of radio resources that can be used to serve the various data flows (recall that they use different frequencies and can be used simultaneously).

1) Direct Transmissions from the BS: We assume that the BS transmits at full power and serves only one user at a time. The service rate of a user at location x if scheduled by the BS is denoted by $C_d(x)$. This rate is a function of the SINR at the receiver and can be well-approximated by Shannon formula:

$$C_d(x) = W_1 \log_2(1 + \frac{P^{\mathrm{BS}}|x|^{-\alpha^{\mathrm{BS}}}}{N_1}), \quad (1)$$

where $|x|$ is the distance from location x to the BS, P^{BS} is the BS's transmission power, α^{BS} is the attenuation exponent and N_1 is the white noise power. This assumption is quite realistic (up to a multiplicative factor) for example in the case of CDMA 1Ev-Do or UMTS/HSDPA systems. We assume there is a direct uplink channel, used for signaling (scheduling, acknowledgments, etc.). In order to simplify the exposition, we will denote with $t(x) = 1/C_d(x)$ the time needed to send a unit of data to location x.

2) Relay Capabilities: The relay channel considered is based on the design principles of 802.11 MAC/PHY (nevertheless, our model is valid for most of other WLAN and WPAN physical layers that are designed on the same principles). It supports variable transmission rates. If a signal, coded for a given rate, is received at an SINR below the corresponding threshold, the packet is lost.

In order to control the interference at the receivers, we use the idea of the RTS/CTS signaling. A node willing to transmit a packet first sends an RTS message, the receiver answers by sending a CTS message. This procedure ensures that no other node will start transmitting in an area around the transmitter and the receiver. For simplicity, we assume that this area consists of all positions at a distance less than D from the transmitter or the receiver. We denote by $I(l)$ the set of links that are disabled by the RTS/CTS procedure[1] initiated by the transmitter and the receiver of link l.

We assume that relay nodes transmit at full power (denoted by P^{RELAY}). P^{RELAY} is assumed to be identical for all nodes. The choice of full power has been extensively justified in the literature on rate-adaptive, multi-hop networks, see e.g. [12], [13], [14].

For a given link we need to choose coding rate C_r as a function of link length l. Packet retransmissions are expensive, and it is important to choose a sufficiently low rate to avoid packet errors. We will choose

$$C_r(l) = W_2 \log_2\left(1 + \frac{P^{\mathrm{RELAY}} l^{-\alpha}}{N_2 + kP^{\mathrm{RELAY}} D^{-\alpha}}\right), \quad (2)$$

where k is a margin factor guaranteeing low packet error rate. This factor is an approximation that quantifies the maximum interference generated by other active relay nodes such that the packet error rate on the link considered remains negligible. We assume it is predefined by a given WLAN's rate adaptation protocol (e.g. $k = 3$).

B. Performance objectives

The users perceive performance through the long-term rate at which their flow is served. In the following, we denote by $\rho(x)$ the long-term rate of a flow whose corresponding user is located at $x \in \mathcal{N}$. We define the cell capacity or throughput by $\rho = \sum_{x\in\mathcal{N}} \rho(x)$.

The goal of the network operator is typically to maximize the revenue of its network [15]. With the revenue in mind, the operator assigns different priority to users at different locations to strike the right balance between the total traffic transmitted and the perceived network quality for users at different locations (fairness). We assume each user is guaranteed a fraction of throughput $p(x)$ that is a function of its distance to the base station. Function $p(x)$ is defined by the operator. This is along the lines of HDR design [16], which assigns different weights to users with different data rates (and the data rate is proportional to the distance from the base station).

We target an allocation of resources maximizing the total throughput, such that each user at distance at x is guaranteed a fraction $p(x)$ of total throughput. In other words, we want to solve:

$$\max \rho, \text{s.t. } \forall x, \rho(x) = \rho p(x).$$

The solution $\rho^\star$ of the above optimization problem is referred to as the *capacity* of the system. Playing with the throughput fractions $p(x)$ allows to tune the trade-off between fairness and efficiency. For example, uniform $p(x)$ ($p(x) = 1$ for all x) corresponds to max-min fairness, whereas having $p(x) = 1/(0.1 + x)$ leads to more efficient but less fair strategy.

We will also make the following assumption on $p(x)$ throughout the paper:

Assumption 1: Function $p(x)$ is non-increasing in $|x|$. This assumption means that we provision less traffic for distant nodes, as serving them costs more, which is perfectly reasonable in a uniformly loaded cellular system (c.f. [16]).

III. Optimal Resources Allocations

Identifying an optimal resource allocation proves extremely difficult in general and we start by introducing a set of simplifying assumptions:

Assumption 2: We assume a fluid queuing model. Also, we assume that the cell is uniformly and heavily loaded: there is a receiver/relay in each small square of size Δx. Each receiver has an unlimited download demand and it is being served with a rate $\rho(x)\Delta x$. We further assume that Δx is sufficiently small that we can approximate our model with the continuous model.

In practice, we can assume that Δx is sufficiently small when the data rates $C_r(y)$ and $C_d(y)$ are do not vary significantly for $y \in (x, x + \Delta x)$, for all values x of interest. A non-uniform traffic can be modelled through the function $p(x)$.

The optimal allocation obtained from the simplified model, using Assumption 2, gives us an upper bound on the capacity. As illustrated later in Section IV, it also provides important insights on how to design optimal resource sharing strategies in real systems (accounting for fading, stochastic queuing dynamics and cells with a finite number of relays).

A. Scheduling and Relay Policies

We now provide a model to describe how radio resources can be shared by active users.

[1] Set $I(l)$ includes link l itself and all the links that share a common node with link l.

1) Scheduling BS resources: The BS shares its power in time between active users. We denote by $\tau(x)$ the proportion of time the BS serves a user at position $x \in \mathcal{N}$. For example, in the Proportional Fair Scheduler of the CDMA 1Ev-DO standard, $\tau(x)$ is inversely proportional to the feasible rate at position x, $C_d(x)$. A feasible scheduling policy is such that:

$$\sum_{x \in \mathcal{N}} \tau(x) \leq 1. \tag{3}$$

2) Relay policies: In the following we denote by $\mathcal{L} \subseteq \mathcal{N}^2$ a set of possible relay links (those whose rate is larger than some minimum). To describe a relay policy, we first define the notion of transmission profile. A profile j is a set of simultaneously active relay links: $j = \{(s_1, d_1), \ldots, (s_p, d_p)\}$. Profile j is feasible if and only if the distance between any pair of nodes (either (s_m, d_n), (s_m, s_n) or (d_m, d_n) for all $m \neq n$) is greater than D (recall that D is the size of RTS/CTS region, defined in Section II). Denote by $\mathcal{J}$ the set of all possible profiles. A relay policy consists of activating the links from profile $j \in \mathcal{J}$ for transmission a proportion of time $\tau_r(j)$. The relay constraint then reads:

$$\sum_{j \in \mathcal{J}} \tau_r(j) \leq 1. \tag{4}$$

A simple example of profile in a 1D cell is a set of equidistant links $j = \{(iD + iD_r, iD + 2iD_r)_{i \in \mathbb{N}}\}$ where D_r is the link length and D is the minimal distance between interfering links.

Unfortunately the number of possible profiles explodes when the number of active users grows, and it then becomes difficult to identify optimal relay policies. Instead, in our theoretical analysis we will use the notion of cliques, see e.g. [9], [17].

Definition 1: A clique is a *maximal* set of links such that two links from this set are not allowed to transmit simultaneously. Here maximal means that a link can not be added to a clique without breaking the previous property.

Denote a clique by Q and the set of all cliques by $\mathcal{Q}$. Let $\tau_r(s,d), ((s,d) \in \mathcal{L})$ be the proportion of time node s sends relay traffic for node d, $\tau_r(s,d) = \sum_{j \in \mathcal{J}:(s,d) \in j} \tau_r(j)$. As demonstrated in [9], any feasible relay policy (e.g. policy that satisfies (4)) has to satisfy the following set of constraints:

$$\sum_{(s,d) \in Q} \tau_r(s,d) \leq 1, \quad \forall Q \in \mathcal{Q}. \tag{5}$$

We will first derive an optimal relay policy satisfying constraints (5). We will then prove that this optimal policy corresponds to an actual policy, i.e., that it also satisfies constraints (4).

Let us define $\phi_d(x) = C_d(x)\tau(x)$ to be the rate of traffic directly sent from the BS to the user at position x, and $\phi_r(s,d) = C_r(|d-s|)\tau_r(s,d)$ to be the rate of traffic sent from the user at position s to the user at position d. Finally denote by $\phi(x)$ the rate at which a user at position x is served. Then a feasible scheduling/relay policy has to satisfy the following flow conservation constraint:

$$\sum_{s:(s,x) \in \mathcal{L}} \phi_r(s,x) + \phi_d(x) = \phi(x) + \sum_{d:(x,d) \in \mathcal{L}} \phi_r(x,d). \tag{6}$$

To summarize characterizing the resource sharing strategy maximizing the weighted system throughput is equivalent to solving the following linear program:

$$\max \ \rho \tag{7}$$

$$\text{s.t.} \sum_{x \in \mathcal{N}} \tau(x) \leq 1, \tag{8}$$

$$\sum_{j \in \mathcal{J}} \tau_r(j) \leq 1, \tag{9}$$

$$\rho p(x) < \phi_d(x) + \sum_s \phi_r(s,x) - \sum_d \phi_r(x,d), \forall x, \tag{10}$$

$$\phi_d(x) = C_d(x)\tau(x), \tag{11}$$

$$\phi_r(s,d) = C_r(|d-s|) \sum_{j \in \mathcal{J}:(s,d) \in j} \tau_r(j), \tag{12}$$

$$\text{over } \tau(x) \geq 0, \tau_r(j) \geq 0, \ \forall x, j. \tag{13}$$

When one considers the relay constraints based on the notion of cliques, the above program is modified replacing constraint (9) by (5), and writing $\phi_r(s,d) = C_r(|d-s|)\tau_r(s,d)$ in (12).

As stated in Assumption 2, in this section we consider heavily loaded cells and we can replace $\sum$ by $\int$ in the problem (7).

3) Existing scheduling and relay policies: We will compare our proposed strategies to two other existing strategies. The first reference strategy is the *direct* policy for which no relaying is allowed [18] ($\tau_r(s,d) = 0$ for all s, d). The second one is the relay-only policy (which we shall call shortly *relay* policy), mimicking policy from [7]. It assumes that only the nodes that do not have any relay in their neighborhood are directly served by the BS ($\tau(d) = 0$ if there exists $s, (s,d) \in \mathcal{L}$).

In the rest of the section, we derive the optimal resource sharing strategy in various scenarios. We first consider the case of the linear, one-dimensional cell where users are located in $[0, R]$, and where the BS is located at 0, with fixed and variable link sizes (Sections III-B and III-C respectively). Later, we will extend the analysis to two-dimensional cells (Section III-D). Notations are summarized in Table I.

B. 1D Cell with Fixed Relay-link Sizes

Consider first the case where the distances between sources and destinations of relay links are fixed and all equal to D_r. The rate of these links is $C_r = C_r(D_r)$ (recall that there are users everywhere). In Section III-C, we will show that relay strategies with fixed, but well chosen, relay link size are almost optimal. We also assume that users at distance $x < D_r$ from the BS may receive relay traffic from the BS[2] (thus, close to the BS, the relay link sizes can be smaller than D_r).

The constraint limiting the BS transmissions is given by:

$$\int_0^R \tau(x) \, \mathrm{d}x \leq 1. \tag{14}$$

Here the cliques are easy to identify: for all $0 < x < R - D - D_r$, the set of links $Q(x) = \{(s, s + D_r) \,|\, s \in [x, x +$

[2] It is reasonable to assume that the BS can also use a (cheap) relay interface for transmission (the results can easily be generalized to the case when this is not true.

$D + D_r]\}$ is a clique, and there are no other cliques in the system[3]. Intuitively, $Q(x)$ is a clique because adding a link to the left or to the right would not interfere with all the links from $Q(x)$. Hence the constraints relative to the cliques are given by:

$$\int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u \leq 1, \quad \forall x \in [0, R - D - D_r]. \tag{15}$$

where[4] $\tau_r(u) = \tau_r([u - D_r]^+, u)$. The flow conservation constraints are:

$$\rho p(x) < C_d(x)\tau(x) + C_r(\tau_r(x) - \tau_r(x + D_r)1_{\{x+D_r \leq R\}}). \tag{16}$$

To simplify the notation, we define $\tau(x) = 0 = \tau_r(x)$ for all $x < 0$ and $x > R$.

We now define a scheduling/relay scheme that will be shown to solve (7). The idea of this scheme is that the weighted cell throughput is strongly limited by users at the cell boundaries, and hence these users should be served by relays only. Formally this scheme is defined as follows.

The MaxRelay scheme: Assume that the cell capacity $\rho^\star$ is known and define X_r by:

$$X_r = \inf(x : \forall y > x, \int_y^{y+D+D_r} \tau'_r(u)\,\mathrm{d}u < 1), \tag{17}$$

where $\tau'_r(x) = \frac{\rho^\star}{C_r} \sum_{i \geq 0} p(x + iD_r)1_{\{x+iD_r \leq R\}}$. Variable $\tau'_r(x)$ may be interpreted as the proportion of time the user at location x should receive relay traffic (from relay the user located at $x - D_r$) destined for itself and all its downstream relay users at locations $x + iD_r$, $i \geq 0$ (as if there were no direct traffic). The MaxRelay scheme is defined by:

$$\tau_r^\star(x) = \begin{cases} \tau'_r(x), & \text{if } x > X_r, \\ \tau'_r(x + D + D_r), & \text{if } x \leq X_r. \end{cases} \tag{18}$$

$$\tau^\star(x) = \begin{cases} 0, & \text{if } x > X_r, \\ t(x)(\rho^\star + C_r(\tau_r^\star(x + D + D_r) - \tau_r^\star(x))), \\ & \text{if } x \leq X_r. \end{cases} \tag{19}$$

Intuitively, (18) means that the relay traffic is carried forward for $x \leq X_r$ and (19) corresponds to (16) for a saturated node. See Figure 1 for illustration.

To complete the definition, the scheme should be such that it uses all the resources of the BS:

$$\int_0^R \tau^\star(x)dx = 1. \tag{20}$$

To prove the optimality of this scheme, we make the following assumption:

Assumption 3: Function $w(x) = t(|x|) - t(|x| - D_r)$ is increasing in $|x|$.

It can be verified by simple calculations of $w(x)$ that the assumption on $w(x)$ is exact when the distance to the BS is not too small (say less than 100m under usual radio propagation models); the MaxRelay scheme proves to be almost optimal even in absence of this assumption.

Theorem 1: Under Assumptions 1, 2 and 3, and assuming fixed size relay link of size D_r, the MaxRelay scheme gives an upper bound on the solution of (7). Moreover, the bound is tight if $p(x) = 1$ for all x.

Proof. Denote by $\rho^\star$ the cell throughput ρ compatible with constraints (14)-(16). It is straightforward to prove that the schedules $\tau^\star$ and $\tau_r^\star$ achieving this maximum are such that the constraints (14) and (16) are saturated. Then we have: $1 = \rho^\star \int_0^R p(x)t(x)\,\mathrm{d}x + C_r \int_0^R t(x)(\tau_r^\star(x+D_r) - \tau_r^\star(x))\,\mathrm{d}x$. Define $t(x) = 0$ if $x < 0$. Now assuming $\rho^\star$ is known, $\tau_r^\star$ is the solution of the following linear program:

$$\begin{array}{lll} \mathbf{LP1}: & \max & \int_0^R \tau_r(x)(t(x) - t(x - D_r))\,\mathrm{d}x \\ & \text{s.t.} & \int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u \leq 1, \\ & & C_r(\tau_r(x) - \tau_r(x + D_r)) \leq \rho^\star p(x), \\ & & \tau_r(x) \geq 0, \forall x,\ \tau_r(x) = 0, \text{for } x > R. \end{array} \tag{21}$$

Assume first that we know the optimal relay scheme $\tau_r^\star(x)$ for all $x > X_r$. Then consider the following linear program:

$$\begin{array}{lll} \mathbf{LP2}: & \max & \int_0^{X_r} \tau_r(x)(t(x) - t(x - D_r))\,\mathrm{d}x \\ & \text{s.t.} & \int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u \leq 1, \\ & & \tau_r(x) = \tau_r^\star(x),\ \text{for } x > X_r. \end{array} \tag{22}$$

If the solution of LP2 satisfies the constraints of LP1, then it will also be the solution of LP1. Denote by $\lambda(x)$ the Lagrange multiplier associated with the first constraint, for $x \in [0, X_r]$. We now identify the term in front of $\tau_r(x)$ in the Langrangian of LP2, $t(x) - t(x - D_r) - \int_{\max(x-D-D_r,0)}^x \lambda(u)\,\mathrm{d}u$ (we write $t(x) = 0, x \leq 0$). This term must be null when τ_r is the solution of LP2. Since the function $t(x) - t(x - D_r)$ is increasing in x we deduce that $\lambda(x) > 0$.

From KKT optimality conditions we conclude that for all $x < X_r$, $\int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u = 1$, which further implies that: $\tau_r(x) = \tau_r(x+D+D_r)$ for all $x < X_r$. The obtained solution satisfies constraints of LP1 (due to the assumption that $p(x) \geq p(x + D + D_r)$), so it must be the solution of LP1. Hence we have proved that LP1 is equivalent to the following linear program.

$$\begin{array}{lll} \mathbf{LP3}: & \max & \int_0^R \tau_r(x)(t(x) - t(x - D_r))\,\mathrm{d}x \\ & \text{s.t.} & \tau_r(x) = \tau_r(x + D + D_r), \forall x < X_r, \\ & & C_r(\tau_r(x) - \tau_r(x + D_r)) \leq \rho^\star p(x), \\ & & \tau_r(x) \geq 0, \forall x,\ \tau_r(x) = 0, \text{for } x > R. \end{array} \tag{23}$$

Now one can easily verify that the solution of LP3 satisfies $\tau_r^\star(x) = \tau'_r(x)$ for all $x \geq X_r$.

Finally we need to show that the MaxRelay scheme can actually be realized for $p(x) = 1$, since constraints relative to the cliques provide an upper bound on the feasible rate region. However, it is easy to see that a simple symmetric schedule, where nodes $D + D_r$ far apart are scheduled at the same time (see the example in Section III-A1) can implement the MaxRelay scheme, which concludes the proof. □

We illustrate the MaxRelay scheme in Figure 1. Immediately, we have:

Corollary 1: In the MaxRelay policy there exist two regions. The first one is for $x < X_r$, and in this region the relay PHY is fully saturated (all cliques are saturated for $x < X_r$,

[3] A proof of this statement in a more general form is given in Lemma 1.

[4] where $[u - D_r]^+ = \max(u - D_r, 0)$

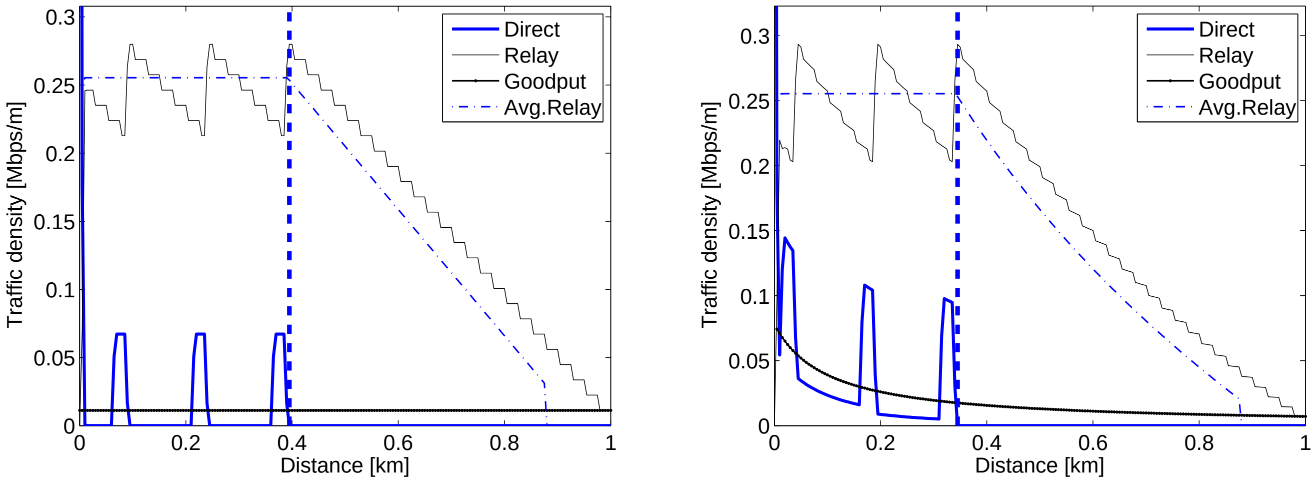

Fig. 1. Examples of optimal scheduling/relay schemes for fixed link lengths ($D_r = 30$m, $D = 100$m). We consider 1D cell of 1km and different throughput fractions $p(x) = 1$ (left) and $p(x) = 1/(0.1 + x)$ (right). The achieved goodput density is $\rho^\star = 12.2$kbps/m. Solid vertical line denotes X_r. Average relay traffic is averaged over each clique $[x, x + D + D_r]$, and it is saturated for $x < X_r$. Periodic peaks of direct traffic correspond to the periodic increase in relay traffic and it occurs at every $D + D_r = 130$m, as predicted by MaxRelay scheme (19).

that is $\int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u = 1$), and some nodes in the region receive both direct and relay traffic. The second region is beyond X_r. For $x > X_r$ the relay PHY is never saturated (no clique is saturated for $x > X_r$, that is $\int_x^{x+D+D_r} \tau_r(u)\,\mathrm{d}u < 1$) and there is no need for direct traffic as it is expensive. This is true regardless of the spatial throughput fraction $p(x)$, although $p(x)$ does influence the values of X_r and $\tau'_r(x)$.
Note finally that the cell capacity $\rho^\star$ is jointly defined with the MaxRelay scheme. It can be easily computed solving (17)-(20).

C. 1D Cell with Variable Relay Link Lengths and Rates

Next we relax the restriction on fixed link lengths. We allow each node to relay over multiple nodes, and we assume that the rate of each relay link depends on its length, as explained in Section II-A2. Our goal is to derive the optimal scheduling strategy and, in particular, the optimal relay routing strategy.

We will proceed as in Section III-B. First, we will identify a region $[0, X_r]$ in which all cliques are saturated and show that in the remaining area $(X_r, R]$ no cliques are saturated. Then we will show that a relay routing using links of a certain fixed length is close to optimal. We will also specify this optimal length.

We cannot theoretically prove the results in this section due to the high complexity of the problem. Instead, we demonstrate them using numerical simulations. We calculate the optimal solution by solving the discrete version of linear program (7)-(13) for 200 equidistant nodes and for different values of network radius R, the exclusion area radius D and throughput fraction $p(x)$. We then compare this optimal result with our proposed heuristic (see Figure 2), and verify the results presented in this subsection.

Before presenting the results, we first need to describe the cliques in the variable link length setting. Let us denote with D_r^{MAX} the maximum allowed relay link length.

Lemma 1: The only cliques that exist in this networks are $Q(x) = \{(s,d) \in [0,R]^2 \,|\, s \le x + D, d \ge x, 0 \le d - s \le D_r^{MAX}\}$, for all $x \in [0, R - D]$.

Proof: We first have to show that $Q(x)$ is a clique, that is, that every two links in $Q(x)$ block each other and that no other such link can be added. It is easy to see that for any two links $(s_1, d_1), (s_2, d_2) \in Q(x)$ we have that $\min(s_1 - d_2, s_2 - d_1) \le D$. Furthermore, we need to verify that if a link (s_1, d_1) does not belong to $Q(x)$ then it is not blocked by all links from $Q(x)$. If $s_1 \le d_1 < x$ then link (s_1, d_1) is not blocked by $(x + D, d), d > x + D$, nor the other way around. If $x + D < s_1 \le d_1$ then link $(s, x), s \le x$ is not blocked by (s_1, d_1), nor the other way around.

Finally, we have to show that there exists no other set $Q' \ne Q(x)$ for all x, which is a clique. Let $x = \min\{d \,|\, (s,d) \in Q'\}$. Then, $x + D \ge \max\{s \,|\, (s,d) \in Q'\}$ because otherwise the two links would not interfere. But the set of points (s,d) that satisfy constraints $d \ge x, s \le x + D, s \le d \le s + D_r$ is exactly $Q(x)$ hence $Q' \subseteq Q(z)$. □

With the above setting, based on numerical and simulations results, we conjecture the following.

Conjecture 1: Under Assumptions 1, 2 and 3, there exists a distance to the base-station $X_r > 0$ such that $\int_{(s,d)\in Q(x)} \tau_r(s,d)\,\mathrm{d}d\,\mathrm{d}s = 1$ for all $x < X_r$. Furthermore, no node after $X_r + D$ receives direct traffic.

The first part of the conjecture has been verified by simulations. One example is depicted in Figure 2 (a). The second part of the conjecture follows immediately. Consider nodes $x, y > X_r + D, y < x$. Both nodes do not belong to any saturated clique. Hence, if $\phi_d(x) > 0$ we can redirect some of the direct traffic to y instead, and forward it from y to x using relay, since it is not saturated. That way we gain some of the BS transmission time, which contradicts with optimality.

Conjecture 2: Consider the model under Assumptions 1, 2 and 3. Let $D_r = \arg\max_l l \cdot C_r(l)$, where $l \cdot C_r(l)$ is the transport capacity of a link of length l, as defined in [19]

(D_r always exists because of the way $C_r(l)$ is chosen). It is approximately optimal for all $d < X_r$ to use as a relay node $s = \min(d - D_r, 0)$. The optimal relay link length for all $d > D_r$ is thus D_r, and is independent of the location of node d.

Again, this conjecture is based on a heuristic verified by simulations. Although we were not able to formally prove it, we provide below some intuitive explanations to justify it.

Using a similar transformation as in **LP1**, we can rewrite the optimization problem (7)-(13) as

LP4 :

$$\max \quad \int_0^R \int_{d-D_r^{MAX}}^{d} \tau_r(s,d) w(d,s)\, \mathrm{d}s\, \mathrm{d}d \tag{24}$$

$$\text{s.t.} \quad \int_{(s,d)\in Q(x)} \tau_r(s,d)\, \mathrm{d}d\, \mathrm{d}s \le 1, \tag{25}$$

$$(\forall x \le R) \int_{x-D_r}^{x} C_r(x-s)\tau_r(s,x)\, \mathrm{d}s - \\ - \int_{x}^{x+D_r} C_r(d-x)\tau_r(x,d)\, \mathrm{d}d \le \rho^\star p(x), \tag{26}$$

$$w(d,s) = C_r(d-s)(t(d)-t(s)), \tag{27}$$

$$\tau_r(s,d) \ge 0. \tag{28}$$

Due to the complex constraints, it is not easy to guess what the solution of this problem is. However, we can see that for the weight associated to link (s,d) in objective function (24) is $w(d,s) = C_r(d-s)(t(d)-t(s))$, and we shall “prefer” links with higher weight.

Let $s(d) = \arg\max_s w(d,s)$, be the relay node with the highest weight with respect to node d. When $d \gg (d - s(d))$, we have $w(d,s) \approx C_r(d-s)t'(d)(d-s)$, and we have $d - s(d) \approx D_r$. However, even when d is of the same order as $d - s(d)$, we verify numerically that $w(d, s(d)) \approx w(d, d - D_r)$. One can interpret the weight $w(d,s)$ as a ratio of time $t(d) - t(s)$ gained on transmitting one bit using the direct link to s instead of transmitting it to d, over the time $1/C_r(d-s)$ needed to relay one bit from s to d. Furthermore, our approximation says that one needs to maximize $(d-s)C_r(d-s)$ which is the rate times the distance. As already mentioned, this is exactly the transport capacity as defined in [19], although in [19] it occurs in a different framework (here, it is a result of a performance ratio between the two physical layers).

Finally, we verify our heuristic numerically. We solve problem **LP4** using linear programming and we compare the optimal routing with our routing heuristic. The results are illustrated in Figure 2. In Figure 2 (b), we see that for $d < X_r$ the optimal routing corresponds well to our heuristic. For $X_r < d < X_r + D$ the optimal link lengths become shorter. This is because the clique $Q(X_r)$ is the last saturated clique, as explained in Proposition 1. Hence for every $s \in Q(X_r)$, it is sufficient to relay data to some node which does not belong to any saturated clique, that is any node $d > X_r + D$. Therefore, link lengths for these nodes tend to be smaller than D_r. Finally, for $d > X_r + D$, relay PHY is not saturated any more hence many routing strategies are possible (including fixed link lengths D_r).

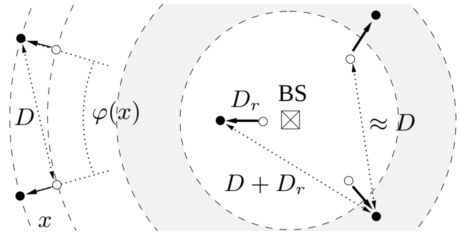

Fig. 3. An illustration of the schedule: the white circle denotes nodes $x < (D + D_r)\sqrt{3}$ where we activate one link at a time and the shaded circle denotes nodes $(D + D_r)\sqrt{3} \le x < D + D_r$ from which we activate two more links, in addition to a link being activated in the inner white circle (as defined in (29)).

We next show that despite of these discrepancies, fixed-length routing with the optimal D_r has a comparable performance to the optimal routing. This is illustrated in Figure 2 (c), where we compare the achieved traffic density ρ of the optimal routing (found by solving **LP4**) and the routing with fixed link lengths D_r for different cell radii R. We see that the error is less than 10%. Furthermore, we verified numerically that the same results hold for different traffic density functions $p(x)$. We verify them for typical parameters for WLAN and WPAN physical layers (numerical details are given in Section V). Finally, the constraints in linear program **LP4** are formed using cliques which represent an upper bound on the actual performance. On the contrary, the performance of fixed-length routing is exact, as explained in Section III-B.

D. 2D Networks

Finally, we consider the case where the cell is a disk of radius R. Again, we first restrict the analysis to the case of fixed relay link sizes, and we discuss variable link length case at the end of the section. As in 1D case, we assume a very large number of users and each user can count on finding a relay in any direction at any distance within the cell's area.

Even with the assumption of fixed link length, deriving the cell capacity is extremely difficult (for example, it proves difficult even to identify cliques). We simplify the problem by the following approximation: we assume only links whose link destinations are on circles of radii $x + k(D + D_r)$, $k \in \mathbb{N}$ may be active at the same time. We next count the maximum number $n_c(x)$ of links that can be simultaneously activated on the circle of radius x. The idea behind the approximation is to map each circle to a node in the 1D case, and to calculate the capacity using the results from Section III-B.

When x is large enough $n_c(x)$ can be well-approximated by $\lfloor 2\pi/\varphi(x) \rfloor$ where the angle $\varphi(x)$ is characterized by $D^2 = x^2 + (x - D_r)^2 - 2x(x - D_r)\cos\varphi(x)$. We can show that this approximation is tight when $x > (D + D_r)/\sqrt{3}$ (for $x = (D + D_r)/\sqrt{3}$, using the approximation we can have 3 simultaneous relay links with receivers at distance x from the BS). Now when $x < (D + D_r)/\sqrt{3}$, one can easily prove that if there is one active relay link with receiver at distance x from the BS, one may add two relay links with receivers at distance $y(x)$ from the BS, where $y(x) = \sqrt{(D + D_r)^2 + x^2 - \sqrt{3}x(D + D_r)}$. All this is illustrated in Figure 3.

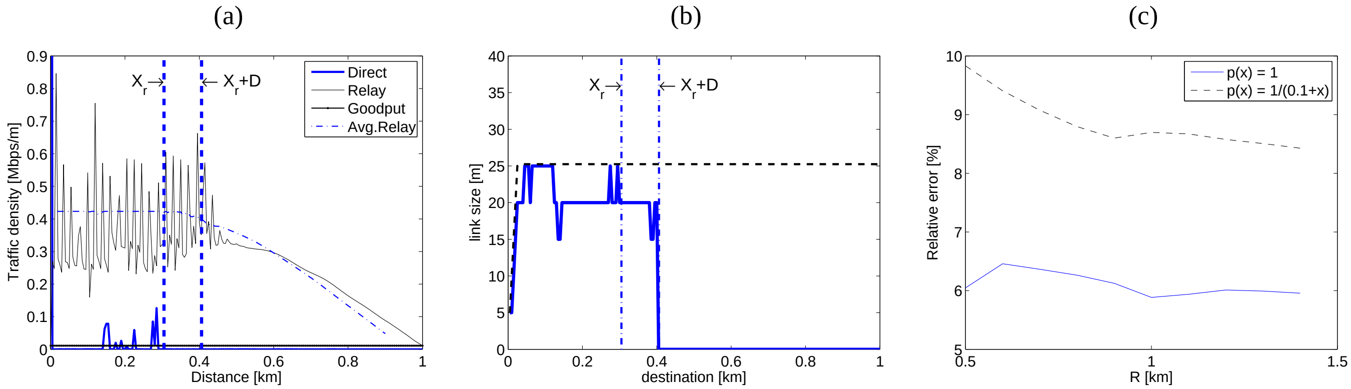

Fig. 2. (a) The optimal traffic distribution (direct vs. relay) for variable link lengths. Relay traffic represents the total relay traffic arriving at node d. The rest is as in Figure 1. (b) The optimal relay link lengths as a function of destination node position. The dashed line represent the optimal routing according to our heuristic. The vertical lines denote X_r and $X_r + D$. (c) Relative error of $\rho^\star$ achieved with fixed routing as compared to the optimal routing for different traffic density $p(x)$.

It is then reasonable to consider the following approximation for $n_c(x)$:

$$n_c(x) = \begin{cases} 1, & \text{if } x < (D + D_r)/\sqrt{3}, \\ \lfloor 2\pi/\varphi(x) \rfloor + 2, & \text{if } \frac{D+D_r}{\sqrt{3}} \le x < (D + D_r), \\ \lfloor 2\pi/\varphi(x) \rfloor, & \text{if } x \ge (D + D_r). \end{cases} \tag{29}$$

It is important to include the two additional activations of nodes in order to densify the schedule around the base-station as this is the most congested area in the relay traffic. Note that in order to make things tractable we make the additional approximation that the two additional activations of nodes on circle $y(x)$ are associated with $\tau_r(y(x))$ and not with $\tau_r(x)$.

Now assume that τ is defined so that $\tau(x)dx$ may be interpreted as the proportion of time the BS is serving all users on the ring between distances x and $x + dx$. Similarly, define τ_r so that $\tau_r(x)dx$ represents the proportion of time users located on the ring between distances x and $x+dx$ from the BS simultaneously receive relay traffic. Note that during this time the BS serves $n_x(x)$ users.

Finally define the throughput fraction p so that $p(x)dx$ is the proportion of traffic received by nodes at distances between x and $x + dx$ (note that it does not imply that the throughput fraction is circular symmetric).

Given the above assumptions, we can find the cell capacity by solving the following maximization problem (instead of (7)-(13))

$$\max\ \rho \tag{30}$$

$$\text{s.t.} \int_0^R \tau(x)\,\mathrm{d}x \le 1, \tag{31}$$

$$\int_x^{\min(R,(x+D+D_r))} \tau_r(u)\,\mathrm{d}u \le 1, \quad (\forall x \in [0, R]) \tag{32}$$

$$\rho p(x) < \tau(x) C_d(x) + C_r(\tau_r(x) n_c(x) - \tau_r(x + D_r) n_c(x + D_r) 1_{\{x+D_r \le R\}}), \quad (\forall x \in [0, R]) \tag{33}$$

$$\text{over } (\tau(x), x \in [0, R]), (\tau_r(x), x \in [0, R]). \tag{34}$$

The 2D MaxRelay scheme: As in 1D cells, we define by $\tau'_r(x)$ as the proportion of time $n_c(x)$ users at distance x should receive relay traffic so as to handle all the traffic to users located at distance $x + iD_r$ from the BS, $i \ge 0$, using relays only, $\tau'_r(x) = \frac{\rho^\star}{C_r n_c(x)} \sum_{i \ge 0} p(x + iD_r) 1_{\{x+iD_r \le R\}}$. Further define X_r as in (17). The 2D MaxRelay scheme is now defined by (18) and:

$$\tau^\star(x) = \begin{cases} 0, & \text{if } x > X_r, \\ t(x)(C_r(n_c(x + D + D_r)\tau^\star_r(x + D + D_r) \\ \quad + \rho^\star p(x) - n_c(x)\tau^\star_r(x))), & \text{if } x \le X_r. \end{cases} \tag{35}$$

Under the approximate scheduling constraints (32) -(33) and the assumption that all links are of the constant length, we can now show that the 2D MaxRelay scheme is capacity optimal. This is formalized in the following theorem:

Theorem 2: Under Assumptions 1, 2 and 3, and assuming fixed-size relay links of size D_r, the 2D MaxRelay scheme, defined by (18) and (35), gives an upper bound to the solution to the optimization problem (30)-(34).

The proof of the theorem is analogous to that of Theorem 1. Consequently, we see that the Corollary 1 holds in 2D case as well.

Finally, we discuss the variable link length case. Since our problem has circular a symmetric structure, we can assume that all nodes on a circle will use links of identical length. Again, we can construct a similar mapping as in the previous case to map the 2D case to the 1D case. Repeating the same type of analysis as in Section III-C, we can verify that the choice of the fixed link length maximizing the transport capacity is approximately optimal in this case as well.

This result also provides an intuitive justification why we can assume D independent of x. Since a region between $[0, X_r]$ has a fully saturated relay traffic, it is likely to expect that the same D and D_r will be optimal throughout this saturated region. Formal verification of this assumption is left for future work.

IV. Routing and Scheduling Protocols

In this section we propose a routing and scheduling algorithms that are derived as heuristics from the results of previous sections and related works (e.g. [11], [7]). These algorithms exploit the benefits that were analyzed in the

(a)

```
$\mathcal{N}_c = \text{BS}, \mathcal{N}_n = \mathcal{N} \setminus \mathcal{N}_c, \mathcal{L}_c = \emptyset,$
for $\mathcal{N}_n \neq \emptyset$
    $s(d) = \arg\max_{s \in \mathcal{N}_c} C_r(|d-s|)|d-s|,$
    $d = \arg\max_{d \in \mathcal{N}_d} C_r(|d-s(d)|)|d-s(d)|,$
    if$|d-s(d)| < D_r^{MAX}$
        $\mathcal{L}_c = \mathcal{L}_c \cup \{(s(d),d)\},$
    end
    $\mathcal{N}_c = \mathcal{N}_c \cup \{d\}, \mathcal{N}_n = \mathcal{N}_n \setminus \{d\}$
end
```

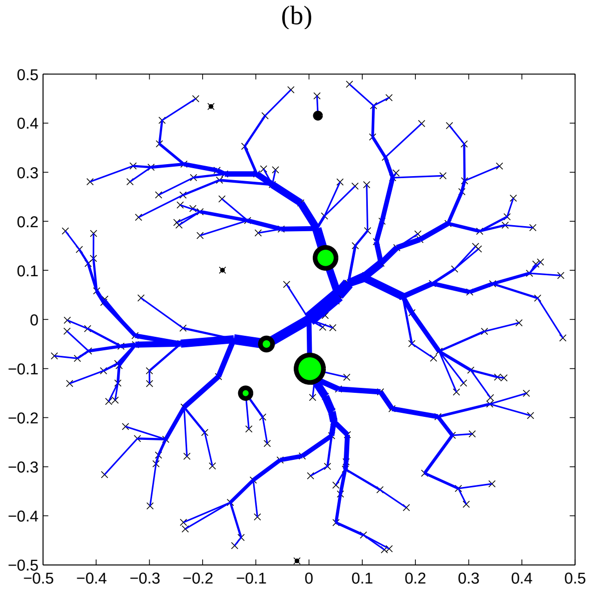

Fig. 4. (a) Relay routing algorithm. $\mathcal{N}_c$ are the connected nodes, $\mathcal{N}_n$ remain to be connected and $\mathcal{L}_c$ are established links. (b) The optimal scheduling and routing scheme for an example of a random topology. We consider 2D cell of radius of 0.5km. We assume $p(x) = 1$ everywhere. The lines represent the relay traffic (the bolder the line is, the higher is the traffic). The circles represent the direct traffic, and the radii correspond to the intensities. The traffic to the disconnected subtrees has to be supplied directly. Also, some of the direct traffic is needed to the areas where the relay interface is congested.

previous sections, and they are simple and can be implemented in a distributed manner.

As discussed in Section III-A, an optimal scheduler maximizes the weighted system throughput subject to constraints (3)-(5)-(6). A way to solve this problem is to implement back-pressure algorithms [11]. However, these algorithms have a high complexity since the number of possible links (relay source-destination pairs) is high.

In order to reduce the complexity, we divide the problem into a routing and a scheduling subproblem. The routing subproblem chooses which relay links shall be used, and the scheduling subproblem solves the above optimization problem constrained on previously selected routes.

A. *Routing Subproblem*

The routing algorithm is based on the results of Section III-C and it is described in Figure 4 (a). Denote by $\mathcal{N}_c$ the set of nodes already connected, by $\mathcal{N}_n = \mathcal{N} \setminus \mathcal{N}_c$ the set of nodes to be connected and by $\mathcal{L}_c$ the set of chosen links. The next node d we connect is chosen so as to maximize the function $C_r(|d-s|)|d-s|$ to any of the already connected nodes $s \in \mathcal{N}_c$, where $C_r(|d-s|)$ is the average rate of link (d,s) over a longer period of time. The candidate source is labeled $s(d)$. If $|d-s(d)| > D_r^{MAX}$, where D_r^{MAX} is the maximum allowed relay link size, it means that node d does not receive relay traffic. Nevertheless, it is put in $\mathcal{N}_c$ as other nodes may connect to it. The routing tree is finally defined by $\mathcal{L}_c$ upon the completion of the above algorithm. The algorithm is illustrated in Figure 4 (b). It can easily be implemented in a distributed manner. Link weights are based on the average link quality (as it is usually the case), rather than on the instantaneous link rates to avoid frequent route oscillations. The variability of the link quality is handled by the scheduling algorithm.

B. *Scheduling Subproblem*

Once the routing algorithm provides us with the set of optimal links $\mathcal{L}_c$, we use a decentralized scheduling algorithm based on the back-pressure principle (as in [11], [7]), restricted on the set of links $\mathcal{L}_c$. We assume that the time $t = 0, 1, \ldots$ is slotted. Let $Q_i(j,t)$ be the number of packets queued at node i for node j in slot t. Node $i = 0$ represents the BS.

The scheduling algorithm shares the resources of the BS and of the relay nodes. In each slot t the BS sends a packet to the destination that maximizes

$$\underset{j \in \mathcal{N}}{\text{argmax}}\, Q_0(j,t) C_d(j,t).$$

To share the relay resources, we use a simple, greedy scheduler that belongs to a class of *maximal scheduling* algorithms [20]. The transmission profile used at a given slot is built iteratively. We start with the following two sets of links $(\mathcal{L}(0), \mathcal{L}_a(0)) = (\mathcal{L}_c, \emptyset)$. At step k, we first identifies a link l such that:

$$l = \underset{(i,j) \in \mathcal{L}(k)}{\text{argmax}}\; Q_i(j,t) C_r((i,j),t),$$

and then we let $\mathcal{L}_a(k+1) = \mathcal{L}_a(k) \cup \{l\}$ and $\mathcal{L}(k+1) = \mathcal{L}(k) \setminus I(l)$, where $I(l)$ is the set of links that interfere with l (including l). We repeat the process until $\mathcal{L}(k)$ is empty set, say for example at step $k = k_f$. The transmission profile to be used is finally $\mathcal{L}_a(k_f)$.

C. Discussion

Our algorithm is similar in spirit to the scheduling algorithm in [7]. It is simpler, suboptimal and can easily be implemented in a distributed manner (e.g. each node sets a back-off timer proportional to the $Q_i(j,t)C_r((i,j),t)$; see for example [21]). One could envisage more sophisticated scheduling algorithms; we let the design and analysis of such algorithms for the future work. Also, note that both the relay and the BS scheduling algorithms are opportunistic in sense that they exploit the information about the quality of channels.

The major benefit of separating the routing from the scheduling subproblem is that the set of candidate links in the scheduling subproblem is significantly reduced. The number of links is given by the routing protocol is $O(N)$ (instead of $O(N^2)$ without routing constraints), and the scheduling protocol converges much faster.

Note that the algorithm proposed in this section is by no means the optimal algorithm for any distribution of users and traffic. In particular, the route construction is oblivious to the traffic demand, as it is assumed that the cell is saturated (hence users demanding traffic are everywhere). The algorithm is an illustration how we can use the findings from Section III to simplify the protocol design and reduce the complexity. As we show in the following section, even such a simple algorithm improves the performance of a saturated cell, as oppose to the conventional direct-only and relay-only approaches. Design of a more robust algorithm is left for future work.

V. Numerical Results

In this section we evaluate the capacity of a single cell network with relays, using the optimal policy derived in the previous sections, and we compare its performance with the direct and relay policies (defined in Section III-A3).

We consider two cases of relay networks. One is WLAN relay and we take typical 802.11 parameters (transmission power 100mW, maximum rate 54Mbps). The other one is WPAN relay and we take next generation 804.15.4a parameters (transmission power 1mW, maximum rate 27Mbps). We assume the BS transmits at 20W and its maximal rate is 10Mbps.

We first look at the optimal resource allocation, as given in Section III-D. This leads to an upper bound of the capacity that would be achieved in a system with no fading, a large number of nodes and that operates a perfect scheduling protocol. Note that it coincides with the one from [18] when only direct traffic is used. In Figure 5 (a) we see that the capacity with the direct policy decreases exponentially (as explained in [18]), whereas it stays constant with the optimal and relay policies. Due to high complexity of stochastic simulations presented below we cannot verify the scaling result using our distributed protocols. Nevertheless, this suggests that relaying can bring significant performance improvements to a cellular network, provided an efficient relay protocol.

Next, we analyze the performance of the distributed routing and scheduling algorithm presented in Section IV. To that end we implement a discrete event simulator that executes the algorithm. We draw a random network of a given radius and of a node density of 250 nodes per km^2. We supply a random traffic to each node in proportion to $p(x)$. In order to keep the network stable we stop injecting any traffic whenever any of the queues reaches over a threshold. We run the algorithm until the average normalized goodput $\rho(x)/p(x)$ to all the nodes approaches the same value ρ. We assume that fading on both direct and relay links is Rayleigh. We implement the direct policy using back-pressure [11] and we implement the relay-only policy using our routing and scheduling algorithm and disabling all direct links (this mimics [7] where only one frequency is available).

The densities of the direct and the relay traffic for a uniform and non-uniform $p(x)$ are shown in Figure 6 (a) and (b). Distance X_r is denoted with a solid vertical line. We see that the ratio of the two traffic is very close to the prediction given in Figure 1. For the distances smaller than X_r the direct traffic is primarily used to reinforce the relay traffic. Vertical dotted lines denote the locations where the maximum of the direct traffic correlates with a drop in relay traffic, similarly as in Figure 1. For the distances larger than X_r the direct traffic significantly drops. Note that the majority of the direct traffic consists of the traffic needed to support the disconnected nodes and subtrees (those that have no direct relay connection to the base station as illustrated in Figure 4 (b)). The remaining direct traffic is used to support the relay traffic, and it is significantly smaller than the goodput, as predicted in Section III.

Figure 6 (c) depicts the average density of queued packets. We see that due to the back-pressure nature of scheduling the queue sizes strictly decrease with the distance from the base-station. Furthermore, queuing information is conveniently used by BS scheduler to discover where the direct traffic should reinforce the relay traffic to keep the queue density curve decreasing.

Figure 5 (b) depicts the cell's capacity as a function of cell size and the routing and scheduling policy and Figure 5 (c) gives a relative performance of our policy over the relay-only one in WLAN and WPAN cases. We see that, although a practical routing and scheduling policies are far from optimal (due to suboptimal scheduling and some nodes not being connected through relays), they still significantly improve the capacity offered by the direct policy and by the relay-only policy (that uses only a single frequency). In particular, for the case of WLAN and WPAN relay networks the performance can be almost doubled by using our distributed routing and scheduling policy as oppose to both the direct and the relay-only policies. Although our policy is not directly comparable with [6], [7] (they use a single frequency for both relays and direct links), we demonstrate that the policy we propose performs significantly better than a naive implementation of the existing single-channel policies on WLAN/WPAN relays.

We also note that there is a discrepancy between the results obtained by the model (Figure 5(a)) and by the simulations (Figure 5(b)). This is due to the node density used in the simulations. Although simulating 250 nodes per square kilometer is at the limit of what we can simulate, it is still far from a saturated network (e.g. a city center of an average town has much higher density).

Small node density limits the amount of spatial reuse in the

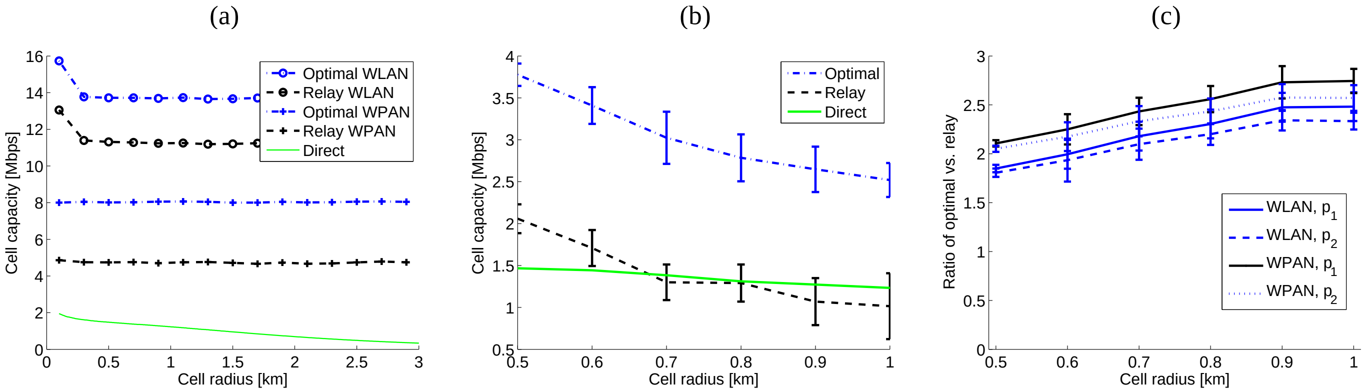

Fig. 5. (a) Capacities of the optimal, relay and direct policies for the case of WLAN and WPAN relay interfaces. (b) The maximum throughput random networks with node density of 250 nodes/km^2 compared to the capacity (for the direct policy the two coincides). (c) Ratio of the capacity achieved by the optimal routing over the capacity achieved by the relay routing for WLAN and WPAN interfaces, for the simulated scenarios ($p_1(x) = 1, p_2(x) = const \times 1/(0.1 + x)$.

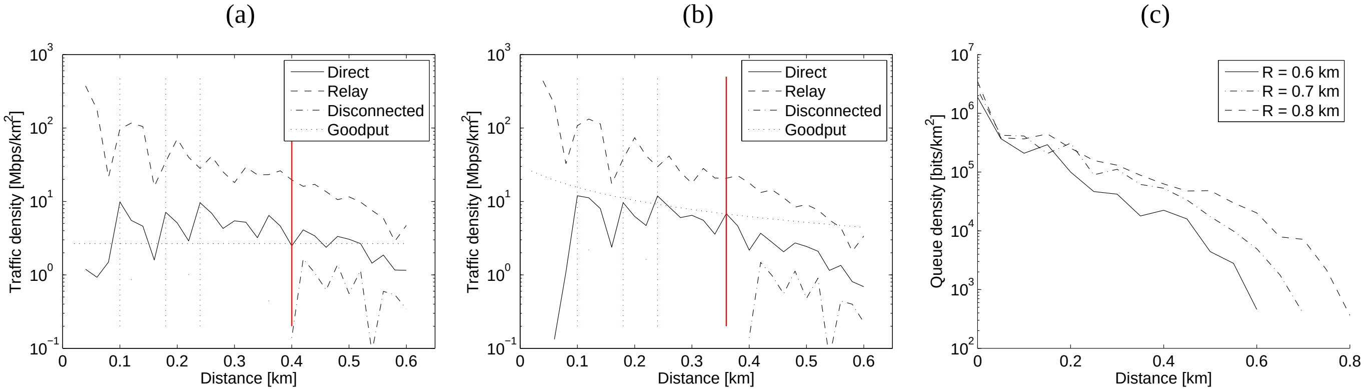

Fig. 6. (a) Average density of total traffic to nodes at a given distance from the base-station, for WLAN, $R = 0.6$km and $p_1(x) = 1$. (b) The same for WLAN, $R = 0.6$km and $p_2(x) = const \times 1/(0.1 + x)$. (c) Average density of queued packets at nodes at a given distance from the base-station, for $p_1(x) = 1$. We divide the x axis into bins of 20m (disjoint rings) and average over all nodes that fall into the bins. In (a) and (b) the *Disconnected* traffic is the total traffic using a direct link to the nodes that cannot use relays since they are disconnected. Vertical solid line represent the approximate X_r. Vertical dotted line represent the peaks of the direct traffic that approximately correspond to the lows of the relay traffic.

system. The model from Section III predicts fully saturated network, hence the optimal spatial reuse. If a possibility for spatial reuse decreases, the total capacity drops sharply, as observed in Figure 5(b).

VI. Conclusions and Future Work

In this paper we have derived an approximately optimal relay (routing and scheduling) policy that maximizes the cell capacity using fixed transmission power. We have shown that it is approximately optimal for relays to use links that maximize the transport capacity. We have also shown that in many cases, a node should receive traffic both from the base station and from a relay, unlike in relay policies proposed by other authors. We have presented a simple algorithm for calculating an upper bound cell capacity (assuming high node density, no fading and the optimal scheduling). Using this bound, we have shown that the cell capacity with relays stays constant with the cell size, as opposed to the capacity of a cell without relays that rapidly decreases with the cell size. We have derived a simple distributed routing and scheduling algorithms based on the findings above. Using extensive simulations we have shown that our optimal strategy largely outperforms other strategies that use direct links only to the nearest node, as proposed in the literature. In future we plan to consider the impact of a bound on a maximum number of relay hops (e.g. due to delay constraints) and possible inefficiencies of a real schedule on the cell capacity. We also intend to evaluate our algorithm on lightly loaded cells.

References

[1] A. Kusuma, L. Andrew, and S. Hanly, "On routing in cdma multihop cellular networks," in *GLOBECOM*, 2004.

[2] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105–2115, 2001.

[3] H. Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu, "UCAN: a unified cellular and ad-hoc network architecture," in *Proceedings of MobiCom '03*, 2003, pp. 353–367.

[4] S. Mengesha, H. Karl, and A. Wolisz, "Capacity increase of multi-hop cellular wlans exploiting data rate adaptation and frequency recycling," in *Proceedings of MedHocNet '04*, 2004.

[5] H.-Y. Wei and R. Gitlin, "Two-hop-relay architecture for next-generation wwan/wlan integration," *IEEE Wireless Communications*, April 2004.

[6] J. Cho and Z. Haas, "On the throughput enhancement of the downstream channel in cellular radio networks through multihop relaying," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 7, pp. 1206–1219, 2004.

[7] H. Viswanathan and S. Mukherjee, "Performance of cellular networks with relays and centralized scheduling," *IEEE trans. on Wireless Communications*, vol. 4, no. 5, September 2005.

[8] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. on Networking*, vol. 14, no. 2, pp. 302–315, 2006.

[9] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE trans. on Information Theory*, vol. 34, no. 5, 1988.

[10] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, January 2005.

[11] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. on Automatic Control*, vol. 37, no. 12, 1992.

[12] B. Radunovic and J.-Y. Le Boudec, "Joint scheduling, power control and routing in symmetric, one-dimensional, multi-hop wireless networks," in *Proc. WiOpt*, 2003.

[13] F. Baccelli, N. Bambos, and C. Chan, "Optimal power, throughput and routing for wireless link arrays," in *Proc. INFOCOM*, Barcelona, Spain, April 2006.

[14] Gjendemsjoe, D. Gesbert, G. Oien, and S. Kiani, "Binary power control for multicell capacity maximization," *Submitted to IEEE Transactions on Wireless Communications*, February 2007.

[15] S. Borst and P. Whiting, "Dynamic rate control algorithms for hdr throughput optimization," in *INFOCOM*, 2001.

[16] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, no. 77, pp. 70–77, July 2000.

[17] T. Nandagopal, T. Kim, X. Gao, and V. Bharghavan, "Achieving MAC layer fairness in wireless packet networks," in *Mobile Computing and Networking*, 2000, pp. 87–98.

[18] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. ACM Mobicom*, San Diego, USA, August 2003.

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

[20] P. Chaporkar, K. Kar, and S. Sarkar, "Throughput guarantees through maximal scheduling in wireless networks," in *Proc. Allerton*, 2005.

[21] U. Akyol, M. Andrews, P. Gupta, J. Hobby, I. Saniee, and A. Stolyar, "Joint scheduling and congestion control in mobile ad-hoc networks," in *INFOCOM*, 2008.

Božidar Radunović is a Researcher in the Networks, Economics and Algorithms group at Microsoft Research, Cambridge. His research interests are in architecture and performance evaluation of computer systems with particular interest in wireless communication, cross-layer design and application of advanced communication techniques in network system design.

Bozidar received his PhD in technical sciences from EPFL, Switzerland, in 2005, and his BSc at the School of Electrical Engineering, University of Belgrade, Serbia, in 1999. He was a PhD student at LCA, EPFL from 2000-2005. Then he did a one year post-doc at TREC, at ENS Paris, in 2006. In 2008 he has been awarded IEEE William R. Bennett Prize Paper Award in the Field of Communications Systems.

Alexandre Proutiere graduated in Mathematics from Ecole Normale Superieure (Paris, France) and received an engineering degree from Telecom Paris Tech in 1998. He received a PhD in Applied Mathematics from Ecole Polytechnique (Palaiseau, France) in 2003. From 1998 to 2000, he worked in the radio communication department at the Ministry of Foreign Affairs in France. He then joined James Roberts' networking research group at France Telecom R&D. From 2007 to 2011, he was a researcher in the systems and Networking laboratory at Microsoft Research, Cambridge, UK. He is now Associate Professor in Automatic Control in the Electrical Engineering School, KTH, Stockholm Sweden. His research interest are in networks, stochastic systems, and learning.

Dr. Proutiere was the recipient in 2009 of the ACM Sigmetrics rising star award for contributions in the analysis and design of distributed control mechanisms in wired and wireless data networks. He received the Best Paper Awards at ACM Sigmetrics / Performance 2004, ACM Mobihoc 2009 and ACM Sigmetrics 2010 conferences. He is currently an Associate Editor for the IEEE transactions on Networking, and for Queueing systems and Applications.