Revisiting Scheduling in Heterogeneous Networks When the Backhaul is Limited

Jagadish Ghimire, *Student Member, IEEE,* and Catherine Rosenberg, *Fellow, IEEE*

***Abstract*—We study the impact of the limited capacity of backhaul links on downlink user scheduling in a heterogeneous network comprising macro base stations and small cells. Assuming a tree topology of the backhaul network, we formulate a *backhaul-aware* global α-fair time-domain user scheduling problem and study it under three different scenarios of backhaul limitations.**

For the scenario where the backhaul links are not the bottleneck, we derive closed-form scheduling solutions to the scheduling problem under certain assumptions. For the scenario where the backhaul links between the macro base station and the small cells are the bottleneck, we show that the global α-fair user scheduling problem can be decomposed into a set of independent local α-fair user scheduling problems. However, unlike the previous case, a local scheduler in this case is not of a unique type but can be of one of three types, depending on the available backhaul capacity. We completely characterize these three types, and also propose a simple heuristic for optimal α-fair scheduling.

When the link between the macro base station and the core network is a potential bottleneck, we show how each base station can still perform a local scheduling as in the previous case as long as there is a master problem that allocates feasible virtual backhaul capacities to each BS. However, computing the optimal virtual capacities is complex and expensive in terms of the amount and frequency of information exchanges. For this scenario, we propose realization-agnostic heuristic schemes that are simple to implement, and perform quite well.

***Index Terms*—Backhaul Limitation, Heterogeneous Cellular Networks, α-Fairness, Proportional Fairness, User Scheduling**

I. Introduction

Heterogeneous Networks (HetNets) comprise a set of low-power base stations (BSs) overlaying the existing macro cellular system [1]. These low power BSs form small cells within the macro cellular coverage area of macro base stations (MBS). They are simply referred to as small cells (SC) throughput this paper. The BSs are connected to the core via some backhaul infrastructure.

The shift from the existing homogeneous structure to a hierarchical heterogeneous architecture offers the potential of a huge improvement in capacity via *network densification* [2] [3].

The complexity introduced due to the addition of an overlapping layer of small cells has led to many studies that revisit the design and operation of important network processes like *User Scheduling* (US), *Resource Allocation* (RA), *User Association* (UA), and *Transmission Coordination* (TC), from a HetNet perspective. It has been shown by a number of works including [4], [5], and [6] that if well designed, these processes can improve the system performance significantly. Most of these studies focus on the wireless access end of the HetNets, and hence there is an implicit assumption that the backhaul infrastructure is not limiting. Such an assumption could be justified in older cellular networks, where the access network (and not the backhaul network) was the bottleneck. In the emerging HetNet architecture, this assumption needs to be reexamined.

In a HetNet scenario, there are two types of backhaul links: the MBS backhaul links (which connect the MBSs to the core), and the SC backhaul links (which connect SCs to the MBS). Network operators see small cell backhauling as an immediate challenge for the successful deployment of HetNets [7], [8]. The ultra-dense deployment of small cells with low average number of users per BS means that the *cost of backhauling* for small cells becomes a significant part of the total Capital Expenditure (CAPEX), in some cases exceeding the cost of the small cell BS equipment [8]. It is thus desirable that the backhauling cost for small cells is kept low. This economic consideration can often limit the capacity of the installed SC backhaul links. For example, a number of cheap solutions are being proposed, including ADSL [9], mesh networks [10], and even non-licensed microwave links [9]. Besides economy, *flexibility* is also a key requirement as there will be numerous SCs added or moved frequently. Fiber or copper infrastructures are often not flexible. The third constraint is *physical*. A small cell might be at an inaccessible street furniture where bringing a fiber link can be infeasible. A low capacity solution like non-line-of-sight (NLOS) wireless backhauling might be the only available option in such a case [7].

MBS backhaul limitations, on the other hand, are less likely to be a concern right now, since MBS backhauling is a small portion of the CAPEX [8], and thus can be well provisioned. However, the future networks are expected to operate with a high number of small cells per macro base-station, with highly efficient wireless links (e.g., using massive MIMO [11]) and on very high bandwidth spectrum (e.g., mmWave [12]). This will translate to a huge increase in traffic load on the backhaul. Moreover, many multi-cell architectures are emerging where signaling for coordination between BSs is done via the backhaul links (e.g., Joint Processing (JP) CoMP [13]), which increases the traffic load on the backhaul links as well as pose more stringent delay requirements. The deployment of cloud-RAN (C-RAN) [14] like architecture is also going to put a lot of pressure on the MBS backhaul. So, it is possible that MBS backhaul limitation might also be a

The material in this paper was presented in part at the IEEE Wireless Communications and Networking Conference, Istanbul, Turkey, April 2014.The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Ave. West, Waterloo, Ontario, Canada ({jghimire,cath}@uwaterloo.ca).

concern for future networks.

Finite capacity of a backhaul link translates into two types of limitations: *1) rate limitation:* the maximum amount of traffic (in bits per seconds) that can be carried via the backhaul link, and 2) *delay limitation:* the delay/jitter incurred by the backhaul link for a given traffic load. These two aspects are inter-related, usually via complex relationships, which are explored using various queuing models. The rate limitation directly affects the total throughput in the HetNet whereas the constraints imposed on delay are key in meeting control signaling deadlines. In this study, we focus only on the rate limitation of the backhaul links, where a backhaul link l has a maximum capacity of C_l Mbps. Note that, limiting the aggregate amount of traffic on a link to a given rate (lower than C_l) can also be used to guarantee a certain level of delay performance on that link.

The exact topology of the backhaul system can have a major impact on the performance. We consider a hierarchical topology of the backhaul links where SC j is connected to the MBS via a backhaul link of capacity C_j and the MBS is connected to the core via a backhaul link of capacity C_{BH}. In other words, for a downlink system, an SC backhaul link has to carry the downlink traffic of its users only whereas the MBS backhaul link has to carry the aggregate traffic of all its users as well as the aggregate traffic from all other SCs in its cell.

The purpose of our study is to understand the impact of backhaul limitations on how user scheduling is to be performed on the downlink of HetNets. Our main message is that finite backhaul links have a fundamental impact on user scheduling, i.e., there is a need for backhaul-aware user schedulers.

We focus on a *macro cellular area* with one macro base station (MBS), and a number of small cells connected to the MBS within a macro cell. We only study the downlink and assume that the resource allocation and the user association scheme are given. For a given *network realization* of channel gains, our objective is to schedule the users at these BSs[1] so as to guarantee fairness. We use the concept of α-fairness, and study user scheduling scheme that guarantees α-fairness in a global sense (i.e., over all users in the considered macro cellular area). By choosing the value of α, an operator can strike the trade-off she wants between fairness and efficiency.

Our contributions can be stated as follows.

1) Our work builds on [5], where Fooladivanda and Rosenberg study the special case of α-fairness where $\alpha = 1$, also called proportional fairness (PF), under unconstraining backhaul capacities. Under this scenario, they have shown that, under some assumptions, the global proportional fair (PF) user scheduling problem decomposes into independent local PF user scheduling problems (one per BS). Additionally, they show that the local PF is equivalent to a local equal-time scheduling scheme. We generalize these results for the general α-fair utility function and in particular derive closed-form expressions for optimal schedules.

2) For the scenario where the MBS backhaul is sufficiently provisioned and hence is not the bottleneck, but where the SC backhaul links have limited capacities, we presented preliminary results in [15] for the special case of $\alpha = 1$. Here, we generalize the results in [15] for any $\alpha > 0$. Our findings for this scenario can be summarized as follows.

- Similar to the scenario of very large SC backhaul capacities, the global problem can be decomposed into independent local problems. The nature of the local α-fair scheduling is different from that of the scenario of very large backhaul capacities. For example, local PF scheduling under backhaul limitations is not always equivalent to the local equal-time scheduling.
- In order to achieve global α-fairness, we show that each small cell j has to schedule its users based on how its backhaul capacity C_j compares to two critical values c_j^* and $C_{j,\alpha}^*$, which are specific to a given network realization. We show that if $C_j \leq c_j^*$ then local α-fair scheduling is equivalent to local *equal-throughput* scheduling, while if $C_j \geq C_{j,\alpha}^*$ then it is equivalent to local α*-fair* scheduling under unconstraining backhaul capacities.
- Using numerical results, we quantify the impact of limited SC backhaul capacity on the system performance. We also propose a heuristic scheduler that is simple to compute and performs very well.

3) For the more general scenario, where the MBS backhaul is also of limited capacity, we perform a detailed analysis of the global scheduling problem, and obtain a number of results. Our findings for this scenario can be summarized as follows.

- We introduce a notion of *virtual backhaul capacity* that allows us to decompose the global problem into per-BS local problems. We present a simple bisection search based algorithm to compute the optimal values of the virtual backhaul capacities. However, these values are realization-dependent and have to be re-computed whenever the network realization changes. In other words, the user schedule at a BS is affected by the channel gains of users in other BSs, which we call the *global realization-dependence* of the optimal solution.
- We present two realization-agnostic heuristics where the *virtual backhaul capacities* are kept fixed all the time, thereby reducing the complexity of the scheduling problem greatly. We quantify the loss in performance due to these schemes and show that they both work well.

The rest of this paper is organized as follows. In Section II, we outline some relevant related work. In Section III, we present the system model. Section IV shows the formulation of the general optimization problem. In Section V, we consider the scenario of unlimited backhaul capacities. In Section VI, we consider the scenario when the MBS backhaul is very large and thus SC backhaul links are the only limitations. Section VII considers the general scenario where the MBS backhaul is also limited. Relevant results are presented in each section. Section VIII concludes the paper.

II. Related Work

Recently, the backhauling aspect of wireless networks has attracted a lot of attention. Its study can be broadly divided

[1]Base Station (BS) refers to both the MBS and the small cells.

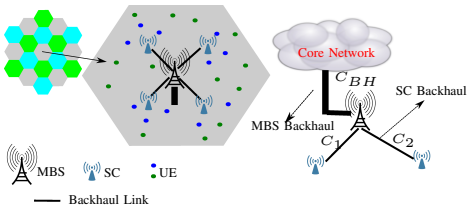


Fig. 1: Our system.

into two types: *provisioning-related* and *impact-related*.

Provisioning-related studies try to characterize the traffic load that a typical cellular deployment imposes on the backhaul network. For example, [16] looks at the LTE-Advanced HetNet deployment and characterizes the traffic load and delay requirements that it can impose in the presence of Joint Transmission based Coordinate Multipoint (CoMP) transmission.

Impact-related studies try to characterize how a limited backhauling can affect the system performance. [17] surveys the impact of limited backhaul on the link level performance due to the reduction in cooperation related capacity gains. Beyond link level performances, backhaul limitations can also impact the user scheduling process in HetNets. There are some studies in the literature that have studied the interplay between backhaul limitation and user scheduling. A number of these works including [18], [19] deal with coordination cluster formation as part of user scheduling decision and they try to make BS clusters so as to reduce the backhaul communication.

Backhaul limitation is not only relevant in multi-cell cooperative transmission. Even in HetNets without BS cooperation, limited backhauls can impact performances due to the delay and/or the rate constraints. Under such limitations, user scheduling decisions have to be made so as to maximize a given system performance by properly utilizing the constrained backhaul resource as well as the precious radio resource. A number of optimization formulations based on *network utility maximization* framework have been proposed in the literature for user scheduling in HetNets (e.g., [5], [4]) for different network-level performance metrics, in the absence of backhaul limitations. Our preliminary work in [15] extends these formulations by considering limited SC backhaul capacities, for proportional fairness. Our current work builds on these work. We take a more general notion of fairness, and consider different scenarios of backhaul limitations. To the best of our knowledge, no such prior work exists.

III. System Model

We consider an OFDM-based cellular network consisting of multiple macro cells. Each macro cell comprises one macro base station (MBS), X small cells (SCs), and N user equipments (UE) (sometimes simply called users), see Fig. 1. We consider each macro cell, with its MBS, SCs, and UEs as a standalone HetNet system. However, we account for interference coming from nearby macro-cells, as we will describe later. We focus on the macro-cell in the middle. 0 represents the MBS, $\mathcal{P} \triangleq \{1, 2, \cdots, X\}$ represents the set of SCs, and $\mathcal{N}$ represents the set of all UEs.

In this study, we consider a *tree* topology of the backhaul network as shown in Fig. 1 where small cell $j \in \mathcal{P}$ is connected to the MBS via a backhaul link of capacity C_j. The capacity of the backhaul link between the MBS and the backbone is given as C_{BH}. Since the major portion of the traffic load is usually on the user plane, we ignore the traffic coming from the control plane[2].

We consider only the *downlink* of the HetNet and assume that all users are active, i.e., there exists a downlink flow from the MBS (source) to each UE (destination). We assume that the users are greedy in throughput and that the BSs have an infinite backlog of packets per UE. The MBS has a transmit power budget of P_{MBS} and each small cell has a transmit power budget of P_{SC}. We assume that each BS transmits all the time with its available transmit power.

A. Subchannel Allocation

The system as a whole uses M' OFDM subchannels and each macro-cell is allocated $M = \frac{M'}{r}$ subchannels, where $r > 1$ is the reuse factor. Thus, a total of M OFDM subchannels are available for the HetNet system under study (i.e., to be used by the MBS in the middle of Fig. 1 and its X SCs).

Different subchannel allocation schemes can be used inside the HetNet, with significant effect on the overall system performance. In this study, we consider a scheme called Orthogonal Deployment (OD) [1], where K subchannels are allocated to the small cells and the remaining $M - K$ subchannels are allocated to the MBS. This exclusive partitioning of subchannels between the MBS and the SCs means that the macro transmissions and SC transmissions do not interfere with each other. In this study, we assume that K is given. The analysis in this work can be applied to other variants, including the partially shared deployment (PSD) and co-channel deployment (CCD) [1].

The following assumptions will allow us to simplify our subsequent formulations: **[A1]** A BS transmits on all the subchannels allocated to it; **[A2]** Power allocated to a given BS is equally divided among all the allocated subchannels; **[A3]** Channels are flat, i.e., the channel gains across different subchannels between a BS and a UE are equal. These assumptions allow us to reduce a time and frequency domain scheduling to pure time domain *single user scheduling* problem, where a BS allocates all of its subchannels to one UE at a given time, as discussed in [20]. However, this means that the channel-dependent scheduling aspect of an OFDM system cannot be exploited in this framework.

A realization $\omega \triangleq \{G_{ji}(\omega)\}_{j \in \{0\} \cup \mathcal{P}, i \in \mathcal{N}}$ represents a set of channel gains between all (BS,UE) pairs. Channel gain $G_{ji}(\omega)$ between BS j and UE i incorporates two random aspects of the network: *1)* the random locations of N users[3], which will result in random path-loss between the BSs and the users, *2)* a random slow fading at each location modeled by a log-normal shadowing of a given standard deviation.

[2]With more complex cooperative communication (like the CoMP) with joint processing, the control plane will also carry a large traffic load in the future.

[3]N (and hence $\mathcal{N}$) can also depend upon ω if we consider a random number of users.

B. Physical interference model and link rates

Let $\gamma_{ji}(\omega)$ be the signal to interference plus noise ratio (SINR) between BS j and UE i on each allocated subchannel for a given realization ω, and for a given P_{MBS} and P_{SC}. For all $j \in \mathcal{P} \cup \{0\}$ and for all $i \in \mathcal{N}$, we have[4]

$$\gamma_{ji}(\omega) = \frac{P_j G_{ji}(\omega)}{N_0 + \mathbf{1}_{\{j\in\mathcal{P}\}} \sum_{k\in\mathcal{P}:k\neq j} P_k G_{ki}(\omega) + \mathbf{1}_{\{j=0\}} I^r_{0i}} \tag{1}$$

where P_j is the power per subchannel for BS j given as

$$P_j = \mathbf{1}_{\{j=0\}} \frac{P_{MBS}}{(M-K)} + \mathbf{1}_{\{j\in\mathcal{P}\}} \frac{P_{SC}}{K}. \tag{2}$$

I^r_{0i} is the interference coming to user i from macro BSs in the nearby macro cells using the same channel resources as MBS 0, based on the reuse factor of r employed among the macro cells. In order to compute this interference, we assume that the nearby HetNets have identical channel allocation scheme (i.e., OD with the same value of K) and transmit power budgets. Interference due to SCs in the nearby cells is often very small. So, for simplicity, we do not consider the interference from SCs in other macro cells, but we do consider interference from SCs in the same cell.

There is a discrete function $\theta(.)$ that maps the SINR $\gamma_{ji}(\omega)$ from BS j to user i to the maximum supportable data rate per subchannel. Then, the maximum supportable rate $R_{ji}(\omega)$ for user i associated to BS j (available only if the UE i is alone in BS j) is given as

$$R_{ji}(\omega) = K_j \theta(\gamma_{ji}(\omega)) \tag{3}$$

where K_j is the number of subchannels allocated to BS j, given by the OD channel allocation scheme, as follows.

$$K_j = K\mathbf{1}_{\{j\in\mathcal{P}\}} + (M-K)\mathbf{1}_{\{j=0\}}, \forall j \in \{0\} \cup \mathcal{P} \tag{4}$$

For a given realization ω, and given backhaul capacities (C_{BH} and $\boldsymbol{C} = [C_1, C_2, \cdots, C_X]$), we assume that the channel allocation parameter K as well as the rate-function $\theta(.)$ are given. In this case, the $R_{ji}(\omega)$'s can be computed a priori as input parameters using (1), (2), (3), and (4). Even though our model assumes that the value of K is given, note that choosing a good value of K is important (and in general not trivial) [20].

C. User Association (UA)

We assume that the user association rule is given, with one UE associating to only one BS. Without loss of generality, we assume that we employ the *Small Cell First (SCF)* user association rule, proposed in [5], with a tunable parameter δ. We choose it as it had the better performance than the other simple UA schemes studied in [5].

Small Cell First (SCF(δ)): UE i associates to small-cell $j \in \mathcal{P}$ if j provides the best per-subchannel SINR $\gamma_{ji}(\omega)$ among all SCs and if this SINR is greater than δ, i.e., if $j = \arg\max_{j'\in\mathcal{P}} \gamma_{j'i}(\omega)$ and if $\gamma_{ji}(\omega) > \delta$. If no such small cell j exists, UE i goes to BS $\tilde{j}$ that provides the best SINR, i.e., $\tilde{j} = \arg\max_{j'\in\{0\}\cup\mathcal{P}} \gamma_{j'i}(\omega)$. Thus, for a given K and ω, this rule with a given value of δ allows us to determine the set of UEs associated to BS j, represented as $A_j(\omega)$. Let $N_j = |A_j(\omega)|$ represent the number of UEs associated to BS j. We assume that the above stated user association scheme guarantees that each UE has a non-zero rate to its BS, i.e., $R_{ji}(\omega) > 0$ for all $j \in \{0\} \cup \mathcal{P}$ and for all $i \in A_j(\omega)$. Note that if $i \notin A_j(\omega)$, then by our definition, $R_{ji}(\omega) = 0$. It is important to note that, even for a fixed value of the UA parameter δ, the sets $A_j(\omega)$ change with the realization.

The backhaul limitations also could have an impact on UA schemes. In this study, we take a simple UA scheme and thus do not consider this impact. Designing backhaul-aware UA scheme is however very important, and we leave it as a future work.

[4]Indicator function $\mathbf{1}_{\{A\}} = 1$ if A is true, 0 otherwise.

IV. Global User Scheduling Problem

We intend to schedule the users so as to guarantee a *global* fairness. This would entail fairness among all users in the entire system, i.e., over multiple cells. However, under our assumptions, the system-level global scheduling can be separated into independent per macro cell scheduling problems. So, in the following, when we mention the *global* problem, we mean the problem at the level of one macro cell, and thus global fairness deals with users within the macro cell under consideration. These users might be associated to the MBS or one of the X SCs.

We use the notion of α-fairness, which was introduced in [21], and has been used often in throughput allocation frameworks usually under Network-Utility Maximization (NUM) formulations [22], [23]. If λ is the throughput offered to a given user, the *utility* corresponding to this allocation is given by $U_\alpha(\lambda) = \frac{\lambda^{1-\alpha}}{1-\alpha}$ if $\alpha > 0, \alpha \neq 1$ and is given by $U_\alpha(\lambda) = \log(\lambda)$ if $\alpha = 1$.

For tractability, we made the assumptions **[A1]**-**[A3]**, which allow us to reduce the scheduling problem to a pure time-domain *single user scheduling* at each BS. Thus, the user scheduling process is completely characterized by $\{\beta_{ji}\}_{j\in\{0\}\cup\mathcal{P}, i\in\mathcal{N}}$, where β_{ji} denotes the fraction of time BS j schedules user i. Then, our global α-fair user scheduling problem corresponds to finding the values of $\{\beta_{ji}\}$'s such that $\sum_{i\in\mathcal{N}} U_\alpha(\lambda_i)$ is maximized, where λ_i is the throughput offered to user i. Of particular interest is the case of $\alpha = 1$, as used in [24] which yields the global proportional fair (PF) scheduling problem.

Formally, the global scheduling problem can be stated as follows: given ω, $\{R_{ji}(\omega)\}$, K, C_{BH}, $\{C_j\}$, find the optimal values of $\{\beta_{ji}\}$ by solving the following.

$$\textbf{[P}(\omega)\textbf{]} \quad \max_{(\lambda_i),(\beta_{ji})} \sum_{i\in\mathcal{N}} U_\alpha(\lambda_i)$$

$$\text{subject to: } \lambda_i = \sum_{j\in\mathcal{P}\cup\{0\}} R_{ji}(\omega)\beta_{ji}, \forall i \in \mathcal{N} \tag{5}$$

$$\sum_{i\in\mathcal{N}} R_{ji}(\omega)\beta_{ji} \leq C_j, \quad \forall j \in \mathcal{P} \tag{6}$$

$$\sum_{j\in\mathcal{P}\cup\{0\}} \sum_{i\in\mathcal{N}} R_{ji}(\omega)\beta_{ji} \leq C_{BH} \tag{7}$$

$$\sum_{i\in\mathcal{N}} \beta_{ji} \leq 1, \qquad \forall j \in \mathcal{P}\cup\{0\} \tag{8}$$

$$0 \leq \beta_{ji} \leq \mathbf{1}_{\{i\in A_j(\omega)\}}, \forall i \in \mathcal{N}, \forall j \in \mathcal{P}\cup\{0\} \tag{9}$$

(5) relates user schedules to throughputs, (6) is the constraint due to finite backhaul capacities at each small-cell. (7) is the constraint due to the limited capacity of the MBS backhaul, which limits the total flows on all BSs. (8) represents the scheduling constraints at each BS. Note that the mention of ω in the parenthesis of the optimization problem name is done to stress on the fact that the given problem is realization-dependent.

We can show that maximizing the sum of the α-fair utility is equivalent to maximizing the following throughput-based metric.

$$\begin{aligned}\overline{T}_\alpha(\{\lambda_i\}_{i\in\mathcal{N}}) &= \left(\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}\lambda_i^{1-\alpha}\right)^{\frac{1}{1-\alpha}}, \alpha>0, \alpha\neq 1 \\ &= \left(\prod_{i\in\mathcal{N}}\lambda_i\right)^{\frac{1}{|\mathcal{N}|}}, \qquad \alpha = 1 \end{aligned} \tag{10}$$

For PF (i.e., $\alpha = 1$), this metric $\overline{T}_1(.)$ represents the geometric mean (GM) of user throughputs. We will refer to $\overline{T}_\alpha(.)$ simply as the *α-mean throughput*.

We identify three scenarios: *Scenario 0, Scenario 1, and Scenario 2*. Scenario 0 is the scenario where the capacities of both the MBS backhaul and the SC backhaul links are large enough not to be bottlenecks , this is true in particular if $C_j > Kr_{max}$, and $C_{BH} > (XK + (M-K))r_{max}$ where $r_{max} = \max_{\gamma\geq 0}\theta(\gamma)$ is the highest value of the rate function). Scenario 1 represents the scenario where the SC backhaul capacities are limited and the MBS backhaul capacity is not constraining. Scenario 2 is the most general scenario where all backhaul links have capacities that are constraining.

Prior work exists for some versions of this problem for $\alpha = 1$ (i.e., global PF). For scenario 0 (i.e., without considering the backhaul limitations (6) and (7)), Fooladivanda and Rosenberg in [5] have shown that the following properties hold.

1) *Decomposability*: The global problem for $\alpha = 1$ can be decoupled into a set of $X+1$ *independent* local PF problems, one per each BS. A local problem for BS j tries to maximize its own local sum of utilities ($\sum_{i\in A_j(\omega)} U_\alpha(\lambda_i)$), without regard to how the scheduling is done in other BSs. A local scheduling solution at BS j depends only on its local information (e.g., values of channel gains of its own users $A_j(\omega)$) which we will refer to as the *local realization dependence*, as opposed to the *global realization dependence* in which schedules in a BS would depend on channel gains in other BSs. *Local realization dependence* is a desirable property.

2) *Equal-time equivalence*: Under the stated assumptions, a local PF scheduling at BS j is equivalent to an *equal-time* scheduling where each user $i \in A_j(\omega)$ is allocated $\frac{1}{|A_j(\omega)|}$ fraction of time.

In our preliminary work in [15], we studied Scenario 1 for $\alpha = 1$ and showed that the above decomposition holds, but the equal-time equivalence does not always hold.

In this paper, we build on these prior works and study the problem under a more general α-fairness objective.

V. Scenario 0: $\{C_j\}$'s and C_{BH} are very large

The following theorem states our results for Scenario 0.

Theorem 1 (Scheduling under Scenario 0): If all backhaul links are very large,

a) *Decomposition*: The global problem **[P(ω)]** can be decoupled into a set of $X+1$ independent local α-fair problems, one per each BS, where the local problem for BS j is

$$[\mathbf{P^j_{Local}}(\omega)]: \max_{\{\beta_{ji}\geq 0\}_{i\in A_j(\omega)}} \sum_{i\in A_j(\omega)} U_\alpha(R_{ji}(\omega)\times\beta_{ji}) \quad \text{s. t.} \sum_{i\in A_j(\omega)}\beta_{ji}\leq 1; \quad \beta_{ji}\geq 0 \tag{11}$$

b) *Closed-form solution*: The following schedule is optimal for the local problem $[\mathbf{P^j_{Local}}(\omega)]$.

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i'\in A_j(\omega)} T_{ji',\alpha}(\omega)}, \forall i \in A_j(\omega), \forall j \in \{0\}\cup\mathcal{P} \tag{12}$$

where $T_{ji,\alpha}(\omega) \triangleq R_{ji}(\omega)^{\frac{1-\alpha}{\alpha}}$.

Proof: The proof is shown in Appendix B. ■

This result means that scheduling is very simple for Scenario 0. The result is the generalization of the known result for $\alpha = 1$, where the local scheduler is the *equal-time* scheduler.

VI. Scenario 1: C_{BH} is very large while $\{C_j\}$'s are not

When C_{BH} is very large, the constraint (7) (*MBS backhaul constraint*) can be removed from the optimization problem $[\mathbf{P}(\omega)]$. Let us call this relaxed problem as $[\mathbf{P}_\infty(\omega)]$. $[\mathbf{P}_\infty(\omega)]$ can be decomposed into a set of local α-fair scheduling problems, one per BS. The local scheduling problem for the MBS is $[\mathbf{P^0_{Local}}(\omega)]$, which is the simple local α-fair scheduling problem without backhaul limitations, defined earlier. SC j should solve the local α-fair scheduling problem with backhaul limitations, shown below.

$$[\mathbf{P^j_{Local}}(\omega,C_j)]: \max_{\{\beta_{ji}\}_{i\in A_j(\omega)}} \sum_{i\in A_j(\omega)} U_\alpha(\beta_{ji}R_{ji}(\omega)) \text{ s.t.}$$

$$\sum_{i\in A_j(\omega)}\beta_{ji}\leq 1, \qquad (\zeta_{j,\omega}) \tag{13}$$

$$\sum_{i\in A_j(\omega)}\beta_{ji}R_{ji}(\omega)\leq C_j, \qquad (\mu_{j,\omega}) \tag{14}$$

$$\beta_{ji}\geq 0, \quad \forall i\in A_j(\omega) \qquad (l_{j,i,\omega}) \tag{15}$$

where $\zeta_{j,\omega}$, $\mu_{j,\omega}$, and $l_{j,i,\omega}$ are the dual variables of the scheduling constraint (13), the total-flow constraint (14), and the non-negativity constraint of user schedules, respectively.

In other words, under Scenario 1, BS j schedules its users independently of other BSs with only its local information (its own backhaul link capacity C_j, and channel gains G_{ji} of its own users only), and thus there is no need for a global entity to assist in the stated decomposition. (12) can be used to obtain

the optimal solution of $[\mathbf{P^0_{Local}}(\omega)]$. In the next subsection, we will derive the solution to the local α-fair scheduling problem $[\mathbf{P^j_{Local}}(\omega,C_j)]$.

A. Local α-fair Scheduling under Backhaul Limitation

If we define the following two critical values of the backhaul capacity for BS j, and realization ω,

$$c_j^*(\omega) \triangleq \frac{|A_j(\omega)|}{\sum_{i\in A_j(\omega)} \frac{1}{R_{ji}(\omega)}}$$
$$C_{j,\alpha}^*(\omega) \triangleq \sum_{i\in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)} \quad (16)$$

then, the nature of the local α-fair scheduling can be characterized as follows.

Theorem 2: The local α-fair scheduling $[\mathbf{P^j_{Local}}(\omega,C_j)]$ can be characterized based on how the backhaul capacity C_j compares to the two critical values. There are three regions:
(a) If $C_j \geq C_{j,\alpha}^*(\omega)$, the scheduler is in Region 1 (which we refer to as backhaul-unlimited (BHU) scheduler), and is given as follows.

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i'\in A_j(\omega)} T_{ji',\alpha}(\omega)}, \quad \forall i \in A_j(\omega) \textbf{ [Region 1]} \quad (17)$$

(b) If $C_j \leq c_{j,\alpha}^*(\omega)$, the scheduler is in Region 2 (which we refer to as local equal-throughput scheduler), and is given as follows.

$$\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \quad \forall i \in A_j(\omega) \quad \textbf{[Region 2]} \quad (18)$$

(c) If $c_j^* < C_j < C_{j,\alpha}^*$, the scheduler is in Region 3. The optimal dual solution is obtained by solving the following equations for $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$.

$$\sum_{i\in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{(\mu_{j,\omega}R_{ji}(\omega)+\zeta_{j,\omega})^{\frac{1}{\alpha}}} = C_j$$
$$\sum_{i\in A_j(\omega)} \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega}R_{ji}(\omega)+\zeta_{j,\omega})^{\frac{1}{\alpha}}} = 1 \quad \textbf{[Region 3]}$$

The primal solution is then given as $\beta_{ji} = T_{ji,\alpha}(\omega) \times (\mu_{j,\omega}R_{ji}(\omega)+\zeta_{j,\omega})^{-\frac{1}{\alpha}}$ for all $i \in A_j(\omega)$.

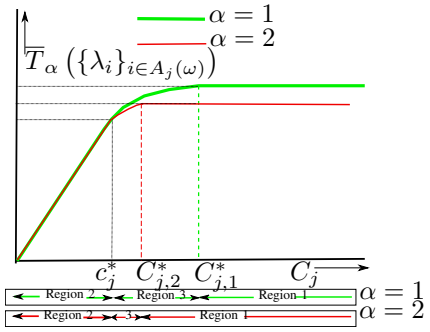
Proof: The proof can be found in Appendix A. ■

Note that the two critical values are realization-dependent which means that any change in the realization would trigger a need to recompute them.

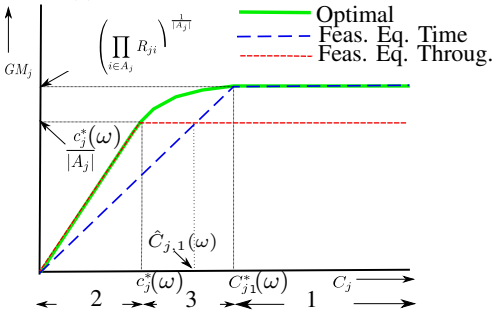
Interpretation of Theorem 2

In Fig. 2a, we show curves that represent the typical shape of the plots of α-mean throughput ($\overline{T}_\alpha(.)$) as a function of the backhaul capacity C_j for a given value of α for one of the small cells $j \in \mathcal{P}$ when the local α-fair scheduling is performed. This figure clearly shows the three scheduling regions (Regions 1, 2 and 3) as a function of the two critical values of the backhaul capacity.

For sufficiently large backhaul capacity $C_j \geq C_{j,\alpha}^*(\omega)$, we are in Region 1. For a very limited backhaul capacity $C_j \leq c_j^*(\omega)$, we are in Region 2. For intermediate values of the backhaul capacity $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, we are in Region 3.



(a) Illustration of Theorem 2



(b) Motivation for the heuristic, $\alpha = 1$

Fig. 2: α-mean throughput versus SC backhaul capacity for a realization

Region 1: For each value of α, there is a critical value of the backhaul capacity $C_{j,\alpha}^*$ such that any more capacity of backhaul link does not translate to a better performance. This is shown as Region 1 in the figure. It is important to note that, for a given set of user rates, this critical value is different for different values of α. Note that in this region, the scheduler is the same as the *backhaul-unlimited* (BHU) scheduler defined for Scenario 0. As an aside, note that $C_{j,\alpha}^*(\omega)$ is also the smallest value of the backhaul capacity C_j for which the backhaul link is no longer a bottleneck on the performance.
Region 2: If $C_j \leq c_j^*(\omega)$, we have $\beta_{ji}R_{ji}(\omega) = \frac{C_j}{|A_j(\omega)|}$ for all $i \in A_j(\omega)$ (from Theorem 2(b)). This is a region where users in a given BS are offered equal throughput $\frac{C_j}{|A_j(\omega)|}$. Thus for $C_j \leq c_j^*(\omega)$, a local *equal-throughput* scheduling is equivalent to the local α-fair scheduling. It is interesting to note that, unlike $C_{j,\alpha}^*(\omega)$, this critical value is independent of α and so is the scheduler. In other words, all α-fair local schedulers operate identically when $C_j \leq c_j^*(\omega)$. In Fig. 2a, they would all have the same Region 2.
Region 3: For $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, neither *local equal-throughput* nor *backhaul-unlimited* α-*fair* scheduling is optimal. The optimal solution to the local α-fair scheduler has to be obtained by computing the solution to the equations in Theorem 2(c). Note that for $\alpha \to \infty$ (i.e., the *max-min* case), Region 3 does not exist.

B. Simple Heuristic

When the scheduling is in Region 1 or Region 2, the variables have closed-form solutions as given in (17) and (18),

and hence are very easy to compute. In Region 1, backhaul-unlimited α-fair scheduling is optimal whereas in Region 2, a local equal-throughput scheduling with throughput of $\frac{C_j}{|A_j(\omega)|}$ is optimal. We do not have closed-form solutions for Region 3, where we need to numerically solve the set of non-linear equations in Theorem 2(c). A scheduler preferably with closed-form solutions for all regions would be desirable.

We propose the following simple heuristic: *take the best of two easy-to-compute feasible schedulers*.

1) The first one is a *feasible* version of the equal-throughput scheduler, i.e., a solution to the local problem with the following constraint $R_{ji}(\omega)\beta_{ji} = R_{ji'}(\omega)\beta_{ji'}$ for all $i, i' \in A_j(\omega)$. The solution to this feasible local equal-throughput scheduling is $\beta_{ji} = \min\{\frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \frac{c_j^*(\omega)}{|A_j(\omega)|R_{ji}(\omega)}\}$ for all $i \in A_j(\omega)$. Note that this scheduler is optimal for Region 2.
2) The second one is a *feasible* (scaled-down) version of the backhaul-unlimited scheduler, i.e., $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} T_{ji',\alpha}(\omega)} k$, where k is a strictly positive scaling constant that corresponds to the largest value less or equal to 1 that guarantees feasibility of the local problem. This problem is solved by $\beta_{ji} = \min\left\{\frac{C_j T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} R_{ji'}(\omega)^{\frac{1}{\alpha}}}, \frac{T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} T_{ji',\alpha}(\omega)}\right\}$ for all $i \in A_j(\omega)$. This scheduler is optimal for Region 1.

The rational behind our heuristic is illustrated in Fig. 2b. This approach results in a much simpler scheduler as compared to the optimal one because of the closed-form scheduling solutions. Of course, we need to verify that this simplification does not result in a significant loss in performance. We will see how this scheme performs in realistic network settings while presenting the numerical results in the next subsection.

Further properties of the local problem: We now present some properties of the local problem that will be used in the analysis of Scenario 2. Let, $f_{j,\omega}(C_j)$ be the optimal value of $[\mathbf{P^j_{Local}}(\omega, C_j)]$. Also, let $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} \triangleq f'_{j,\omega}(C_j)$ be the rate at which the optimal value changes with C_j. Then, we can show that the following holds.

Lemma 1: The rate of change of $f_{j,\omega}(C_j)$ with respect to the backhaul capacity C_j is given as follows:

$$\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} = \begin{cases} \left(\frac{|A_j(\omega)|}{C_j}\right)^{\alpha} & \text{if } C_j \leq c_j^*(\omega) \\ \mu_{j,\omega}^*(C_j) & \text{if } c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega) \\ 0 & \text{if } C_j \geq C_{j,\alpha}^*(\omega) \end{cases} \tag{19}$$

where $\mu_{j,\omega}^*(C_j)$ is the optimal value of the dual variable $\mu_{j,\omega}$ for backhaul capacity C_j.

Proof: Please see Appendix C. ∎

Also, note that $f_{j,\omega}(C_j)$ is a concave, non-decreasing function of C_j in $(0, \infty)$. In particular, $f_{j,\omega}(C_j)$ is strictly increasing in $(0, C_{j,\alpha}^*(\omega)]$. $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} = f'_{j,\omega}(C_j)$ is a strictly decreasing function of C_j in $(0, C_{j,\alpha}^*(\omega)]$.

C. Numerical Results

We consider a hexagonal HetNet deployment area with each side equal to $500/\sqrt{3}$ m, which corresponds to the scenario of an inter-site distance (ISD) of $500m$ (urban setting). The centrally placed MBS is overlaid with $X = 4$ *symmetrically placed* small cells ($j = 1, 2, 3, 4$) at a distance of $d = 178$ m. from the MBS. An MBS transmit power budget P_{MBS} of 46 dBm and an SC transmit power budget P_{SC} of 30 dBm are considered. The overall system has $M' = 99$ subchannels and the reuse factor of $r = 3$. Hence there are $M = 33$ subchannels available to each macro-cell, out of which K subchannels are allocated to each small cell and the remaining $M - K$ subchannels are allocated to the MBS. The interference from the outer macro-cells is calculated by considering 18 identical macro cells around the given macro cell, and by assuming that identical channel splitting (K) is employed in the interfering macro-cells. We only consider the interference the 4 small cells create for each other.

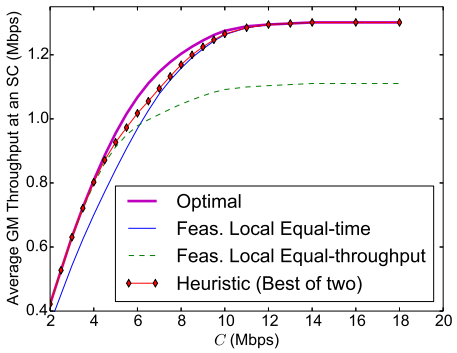
We use the distance-based path-loss model recommended by 3GPP [25], with MBS-UE path-loss at distance $d \geq 35m$ given as $128.1 + 37.6 \log_{10}(d/1000)$ and an SC-UE path-loss at distance $d \geq 10m$ given as $140.7 + 36.7 \log_{10}(d/1000)$ in dB. The channel gains G_{ji} are obtained by further applying a log-normal shadowing of 8 dB standard deviation. A random realization ω corresponds to a realization of channel gains for a random instance of uniformly deployed N equal to 30 user positions and randomly generated shadowing coefficients.

The rate function $\theta(.)$ is taken as the 15-rate MCS available in LTE, as shown in Table III of [20]. The table shows the per-subcarrier efficiency e_l in terms of bits per symbol for a given threshold SNR. The actual per-subchannel link rate $f(\gamma_{ji})$ can then be calculated as $e_l \frac{n_{sc} \cdot n_{sym}}{T_{subframe}}$ if γ_{ji} is between l^{th} and $(l+1)^{th}$ SNR threshold. $n_{sc} = 12$, $n_{sym} = 14$ and $T_{subframe} = 1ms$ are respectively the number of subcarriers in one subchannel, the number of OFDM symbols in one subframe and the duration of a subframe. We take $N_0 = -112.45dBm$ as the noise power per subchannel (i.e., a noise of $-174dBm/Hz$ with a noise figure of 9 dB).

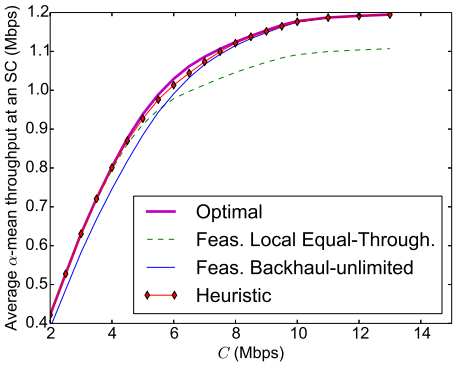
We consider scenarios where the small cells are identical, i.e., they all have the same backhaul capacities C_j equal to C. Also, recall that in this scenario, C_{BH} is sufficiently large.

We study 100 random realizations $\omega \in \Omega$ of user positions. The average of the α-mean throughput $\overline{T}_{\alpha}(.)$ over these realizations is the metric for comparison of the different schemes (optimal and sub-optimal).

1) PF Scheduling ($\alpha = 1$) : In Fig. 3a, we plot the average GM throughput (which is the α-mean throughput for $\alpha = 1$) of users in a given SC as a function of SC backhaul capacity C for $K = 15$, $\delta = 6.6dB$. The impact of limited SC backhaul capacity on the throughput performance is significant. Let us concentrate on the optimal scheduling scheme. We can see that after a certain point, increasing capacity C does not translate to a significant improvement on the throughput performance. This is expected due to the concavity of $f_{j,\omega}(C_j)$, and the fact that $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j}$ is equal to 0 for $C_j \geq C_{j,\alpha}^*(\omega)$. Also, it can be observed that, for this particular scenario, there is a value of C (about 12 Mbps, shown by the vertical dashed line) after which



(a) Average GM throughput versus C, Local PF



(b) Average α-mean throughput versus C, $\alpha = 2$

Fig. 3: Comparison of the optimal and the sub-optimal local α-fair schedulers

there is effectively no improvement in system performance as we increase the backhaul capacity. This value of C can be considered as the *sufficient capacity* of the SC backhaul link to guarantee that the SC backhaul link is no longer a bottleneck to system performance.

The figure also illustrates that using either a local equal-throughput or a local equal-time scheduling regardless of the backhaul capacity can result in a significant loss in performance. The same figure shows that our heuristic, where a BS chooses the best of local feasible equal-throughput and local feasible equal-time for each realization, works remarkably well. This scheduler achieves a performance very close to optimal and yet is quite simple to compute. This heuristic can thus be seen as good *backhaul-aware* local PF scheduler.

2) $\alpha = 2$: In Fig. 3b, we plot results for $\alpha = 2$. This value of α maximizes the harmonic mean of user throughputs and is often called the *minimum potential-delay scheduling*. The plot shows similar results, in particular, it shows that our proposed heuristic is a good approximation of the local α-fair scheduler.

VII. Scenario 2: $\{C_j\}$'s and C_{BH} are not very large

A. Optimal Scheduler

Similar to Scenario 1, it would be desirable to decompose the global problem for Scenario 2 into independent local problems. However, unlike Scenario 1, decomposing the global problem into local problems is not straightforward, mainly due to the coupling constraint (7) (the MBS backhaul constraint). Indeed, allowing each BS j to independently schedule based on its own local problem ($[\mathbf{P^j_{Local}}(\omega,C_j)]$) could lead to violation of the MBS backhaul constraint (7). Thus, in order to obtain a decomposition that is feasible, we need to guarantee that the local problems do not violate the MBS backhaul constraint. This can be accomplished by defining the notion of *virtual backhaul capacities* $\tilde{C}_j$ for each BS $j \in \{0\} \cup \mathcal{P}$ where $\tilde{C}_j$ is used by the local scheduler at BS j as the actual available local capacity (as opposed to C_j). The vector $\tilde{\boldsymbol{C}} = [\tilde{C}_0, \cdots, \tilde{C}_j, \cdots, \tilde{C}_X]$ is considered to be feasible if it satisfies the following conditions.

$$\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j \le C_{BH}; \quad \tilde{C}_j \le C_j, \quad \forall j \in \mathcal{P} \tag{20}$$

Given a feasible vector $\tilde{\boldsymbol{C}}$, if the local scheduler at BS j solves $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$ with $\tilde{C}_j$ as its local backhaul constraint without regard to other local problems, the end result is a feasible solution to the global problem $[\mathbf{P}(\omega)]$. Hence, as long as a *master* problem can provide one such feasible $\tilde{\boldsymbol{C}}$, the solutions due to the local schedulers would yield a feasible solution to the global problem, thereby yielding a *feasible* decomposition. Our first goal is to find an *optimal* decomposition, i.e., one that would yield the optimal solution of problem $\mathbf{P}[\omega)]$. Solving the following master problem will provide the values of $\tilde{C}_j$ corresponding to the optimal decomposition.

$$\max_{\tilde{\boldsymbol{C}}\ge 0} \sum_{j\in\{0\}\cup\mathcal{P}} f_{j,\omega}(\tilde{C}_j) \text{ s.t. } (20) \tag{21}$$

Recall that $f_{j,\omega}(\tilde{C}_j)$ is the value of the local problem $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$ at BS j where $\tilde{C}_j$ is the backhaul capacity.

Without loss of generality, we will assume that the MBS solves this master problem. In our tree topology, it is indeed the most natural place to compute the solution to the master problem. This formulation can be seen as a *two-level* problem in which small cells report their user channel gain information ($\{G_{ji}(\omega)\}_{i\in A_j(\omega)}$) to the MBS which computes and reports back to them the optimal values of the virtual capacities $\tilde{C}_j$. BS j can then perform the local scheduling by considering the reported $\tilde{C}_j$ as the available backhaul capacity (as opposed to C_j), i.e., solving $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$. This two-level decomposition solves the global problem optimally, and can be seen as an alternative formulation to $[\mathbf{P}(\omega)]$. Note that presenting the original global problem as a two step problem does not simplify its computational complexity but allows us to propose a good heuristic later.

Alternatively, a distributed approach could be used where the small-cells would report the subgradients and the MBS would update them with the value of Λ. Such process would eventually converge to the optimal solution. This way, there is no need to collect the channel gains, but now the overhead would be on the exchange of Λ, and one gradient per BS at each step until convergence. The efficiency would depend on how fast the process converges.

Alternative Dual-based Formulation

We can rewrite the master problem as: $\max_{\tilde{\boldsymbol{C}}\geq 0} \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ s.t. $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j \leq C_{BH}$, where $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j) = \min\{f_{j,\omega}(\tilde{C}_j), f_{j,\omega}(C_j)\}$ for all $j \in \{0\}\cup\mathcal{P}$, and where C_0, which is not a physical constraint but is here for consistency, is equal to a large value (e.g., greater than C_{BH}). This modified problem has only one dual variable, which we call Λ. Then, solving the following dual problem is equivalent to solving the master problem,

$$\min_{\Lambda\geq 0} \quad \max_{\tilde{\boldsymbol{C}}\geq 0} \quad L(\tilde{\boldsymbol{C}};\Lambda) \tag{22}$$

where $L(\tilde{\boldsymbol{C}};\Lambda) = \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j) - \Lambda\left(\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j - C_{BH}\right)$ is the Lagrangian function.

The following result allows us to obtain the solution to the dual problem.

Theorem 3:

$$\Lambda^*(\omega) = \min\left\{\Lambda\geq 0 : \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda) \leq C_{BH}\right\}$$

is the optimal value of Λ in problem (22), where $\tilde{C}^D_{j,\omega,C_j}(\Lambda) \triangleq \min\{f'^{(-1)}_{j,\omega}(\Lambda), C_j\}$ for all $j \in \{0\}\cup\mathcal{P}$ is a mapping from dual variable Λ to primal variable $\tilde{C}_j$, $f'^{(-1)}_{j,\omega}(\Lambda)$ is the inverse mapping of $f'_{j,\omega}(C_j)$ defined in (19), and is given as follows:

$$f'^{(-1)}_{j,\omega}(\Lambda) = \begin{cases} \frac{|A_j(\omega)|}{\Lambda^{\frac{1}{\alpha}}} & \Lambda \geq \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^{\alpha} \\ \mu^{*(-1)}_{j,\omega}(\Lambda) & 0 < \Lambda < \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^{\alpha} \\ C^*_{j,\alpha}(\omega) & \Lambda = 0 \end{cases} \tag{23}$$

where $\mu^{*(-1)}_{j,\omega}(\Lambda)$ is the inverse of $\mu^*_{j,\omega}(C_j)$ which is the dual variable of the local problem $[\mathbf{P^j_{Local}}(\omega,C_j)]$ as defined earlier, and $(c^*_j(\omega), C^*_{j,\alpha}(\omega))$ are the critical values defined in (16).

Proof: Please see Appendix D. ∎

Even though computationally similar, there is a benefit of looking at the dual version as opposed to the primal problem: we can find a mapping from the dual variable to the virtual backhaul capacity $\tilde{C}_j$ allowing us to express the primal optimal solutions based on the optimal value of Λ.

This problem can be solved for one scalar value of Λ by employing a simple *bisection-search* for the smallest feasible value of Λ. This is because, the $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ are non-decreasing as we decrease Λ, as shown in Appendix D (and hence $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda)$ is non-decreasing as we decrease Λ).

The details of the bisection search algorithm is presented in Algorithm 1.

In the dual framework, we can thus view the global optimization as follows: The MBS computes the optimal dual variable $\Lambda^*(\omega)$ and sends this value to the SCs. Each SC computes its virtual backhaul capacity $\tilde{C}_j = \tilde{C}^D_{j,\omega,C_j}(\Lambda^*(\omega))$ for the given dual variable and then performs its local scheduling using this computed value. Besides this concise representation of the optimal solution, the dual formulation also serves as the basis for a very good heuristic that we will discuss later.

Algorithm 1 Compute Optimal Dual Variable $\Lambda^*(\omega)$

Input: $\omega, \{\tilde{C}^D_{j,\omega,C_j}(.)\}_j, C_{BH}$
Define: $g(\Lambda) \triangleq \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda)$
Begin:
$(\Lambda_{Max}, \Lambda_{Min}) \leftarrow (L, 0)$ ▷ *L :A sufficiently large number*
while $|\Lambda_{mid} - \Lambda_{Max}| < \epsilon$ **do** ▷ *ϵ :A small positive number*
$\Lambda_{Mid} \leftarrow \frac{\Lambda_{Max}+\Lambda_{Min}}{2}$
if $g(\Lambda_{Mid}) \leq C_{BH}$ **then** $\Lambda_{Max} \leftarrow \Lambda_{Mid}$
else $\Lambda_{Min} \leftarrow \Lambda_{Mid}$
end if
end while
Return Λ_{Max}

B. Complexity and Overhead versus Performance Trade-off: Heuristic Schemes

The optimal values of $\tilde{C}_j$ (either computed using the primal master problem or by using the dual version) are **global realization-dependent**. By global realization-dependence, we mean that the values of $\tilde{C}_j$ change with a change in the network realization. Such a change in realization could be due to various factors: mobility, change in channel gains, user arrival or departure etc. There are at least two aspects of the optimal scheduler that are undesirable:

1) **Computational Complexity:** Computationally, the master problem is as complex as the global problem $[\mathbf{P}(\omega)]$ which is a convex optimization problem of size $\Theta(XN)$ with $\Theta(XN)$ constraints. The complexity of an interior-point method for solving a convex optimization problem is known to be polynomial on the problem size [26]. In our case, the problem size (the number of variables or constraints) increases linearly with the number of users N. Thus, for medium to large values of N, quick computation of the optimal $\tilde{C}_j$ can be a challenge.

2) **Amount and frequency of information exchange (overhead)**: The master problem needs the information of the channel gains from all users in all BSs. The optimal problem (either the primal or the dual version) is complex as it requires re-computation of the master problem each time a realization changes.

Note that other key parameters such as the resource allocation parameter K or the UA parameter δ can also change with time. However, the time-scale at which these parameters change is usually much larger than the time-scale at which the realizations change. So, in the remainder of this section, we assume that K and δ are fixed and do not change with time.

It would be desirable to have a scheme that overcomes the aforementioned issues by: 1) having a simpler master problem (with reduced problem size), and 2) requiring less overhead (i.e., less amount of and less frequent information exchange between the MBS and the SCs) for solving the master problem. These simplifications come at the expense of some loss in throughput performance. Finding the right amount of trade-off between the throughput performance and the complexity/overhead is important.

We can define a class of *realization-agnostic* schemes where the virtual backhaul capacities are kept fixed all the time, even when the network realization changes, with changing channel gains as well as changing number of users in the system. Given

these fixed values of the virtual backhaul capacities, the global problem can then be decomposed into per-BS local problems, thereby not requiring any information exchange between the MBS and the SCs.

We will take two approaches to choosing the realization-independent values of the virtual backhaul capacities:

Average virtual backhaul capacity based approach:
We could generate offline a set of realizations in Ω' where each realization $\omega \in \Omega'$ has a random number of users N. We can then compute the optimal values of the virtual backhaul capacities for each of these realizations. We can then take the average of these virtual capacities as our fixed values, i.e.,

$$\overline{C}_j = \frac{1}{|\Omega'|} \sum_{\omega \in \Omega'} \tilde{C}_j^*(\omega), \forall j \in \mathcal{P} \cup \{0\}$$

Dual-based approach:
We can replace the master problem by an approximate problem that is realization-independent. For example, in the dual version of the master problem, we can replace the realization-dependent mapping $\tilde{C}^D_{j,\omega,C_j}(.)$ by a realization-independent mapping $\overline{C}^D_{j,C_j}(.)$. Below, we derive one such mapping.

From Theorem 3, we know that in the interval $[\left(\frac{N_j(\omega)}{c_j^*(\omega)}\right)^\alpha, \infty)$, $C^D_{j,\omega,C_j}(\Lambda)$ is a non-increasing function of Λ with dependence on $N_j(\omega)$, and is equal to $\min\{C_j, \frac{N_j(\omega)}{\Lambda^{\frac{1}{\alpha}}}\}$. Also, $C^D_{j,\omega,C_j}(\Lambda)$ is a non-linear non-increasing function in the interval of $[0, \left(\frac{N_j(\omega)}{c_j^*(\omega)}\right)^\alpha]$ (that depends on the actual rates $\{R_{ji}(\omega)\}$), decreasing from $\min\{C_j, C^*_{j,\alpha}(\omega)\}$ for $\Lambda = 0$ to $\min\{C_j, c_j^*(\omega)\}$ for $\Lambda = \left(\frac{N_j(\omega)}{c_j^*(\omega)}\right)^\alpha$. If we replace the instantaneous values of $c_j^*(\omega)$, $C^*_{j,\alpha}(\omega)$, and $N_j(\omega)$ by the average values of these quantities, we could achieve our goal of replacing $C^D_{j,\omega,C_j}(\Lambda)$ by functions of Λ that do not depend on the realization, as follows:

Given channel allocation parameter K, UA parameter δ, and a set of realizations Ω, we can compute the *average* values of $C^*_{j,\alpha}(\omega)$, $N_j(\omega)$, and $c_j^*(\omega)$: $\overline{N}_j \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} N_j(\omega)$, $\overline{C}^*_{j,\alpha} \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} C^*_{j,\alpha}(\omega)$, and $\overline{c}^*_j \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} c^*_j(\omega)$. We can then use the following simple relationships between the (approximate) dual variable Λ and the primal variables $\tilde{C}_j$: $\overline{C}^D_{j,C_j}(\Lambda) = \min\{\overline{f}'^{(-1)}_j(\Lambda), C_j\}$ where

$$\overline{f}'^{(-1)}_j(\Lambda) = \begin{cases} \frac{\overline{N}_j}{\Lambda^{\frac{1}{\alpha}}}, & \Lambda \geq \left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha \\ \left(\overline{C}^*_{j,\alpha} - \Lambda \times \Delta_j\right), & \Lambda < \left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha \end{cases}$$

and $\Delta_j \triangleq \left(\frac{\overline{C}^*_{j,\alpha} - \overline{c}^*_j}{\left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha}\right)$.

The dual-based scheme works as follows: The small cells report the measurements on the average values of $(\overline{c}^*_j, \overline{C}^*_{j,\alpha}, \overline{N}_j)$. With these values, the MBS uses the bisection-search algorithm in Algorithm 1 to compute the realization-agnostic values of the virtual backhaul capacities which it sends to the SCs. These values are then kept fixed.

Remark 1: Note that the dual-based heuristic can be implemented easily as an online algorithm (with no offline tuning required). This can be done by each BS learning the required averages, and reporting these averages once the measurements converge.

C. Numerical Results

We study how the realization-agnostic schemes work over a set of 500 realizations Ω', where each realization $\omega \in \Omega'$ has a number of users chosen uniformly at random in the interval $[10, 30]$. The users are distributed uniformly at random in the deployment area. Note that, in Section VI-C, we considered a set of realizations Ω with a fixed number of users ($N = 30$). But, in this section, we consider realizations with different number of users. This setup encompasses a large set of random realizations in a dynamic network with varying number of users and thus allows us to see if the realization-agnostic scheme works well in a dynamic context. Other than this, we take the same physical layer and network level parameters and setup as in Section VI-C.

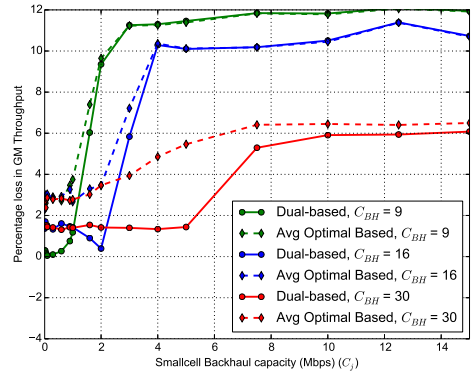
In Fig. 4a, we consider the case of proportional fairness (PF) ($\alpha = 1$) with three different values of C_{BH}, i.e., 9, 16, and 30 Mbps, for $K = 15$ and $\delta = 6.6dB$. We present the performance of the sub-optimal schemes in terms of the loss in α-mean throughput performance incurred due to these schemes with respect to the optimal one, for different values of the backhaul capacities. Let $\chi^s(\omega)$ be the α-mean throughput for realization ω for scheme s. Then, the average loss in α-mean throughput for scheme s over the set of realizations Ω' is given as $100 \times \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} \frac{\chi^{(Opt)}(\omega) - \chi^{(s)}(\omega)}{\chi^{(Opt)}(\omega)}$ where $\chi^{(Opt)}(\omega)$ is the α-mean throughput of the optimal scheme for realization ω.

Observation (*Realization-agnostic schemes work well for* $\alpha = 1$): The results show that the price of using a realization-agnostic scheme is less than 12% for C_{BH} small and decreases when C_{BH} increases. A degradation of less than 12% is a reasonable price to pay, especially since the optimal scheme would be much more complex, and would require a lot of information exchange and a frequent global computation of the optimal solutions. A realization-agnostic scheme, on the other hand, yields independent scheduling at each BS.

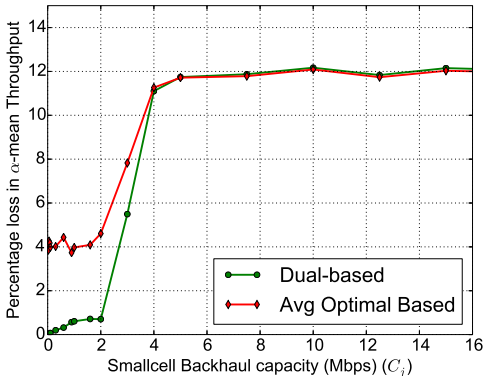
In Fig. 4b, we show similar results for $\alpha = 2$ for $C_{BH} = 16Mbps$. This shows the effectiveness of our heuristic schemes for another value of α.

VIII. Conclusion

In this paper, we study the impact of limited backhaul capacity on user scheduling in a heterogeneous network with a macro base station (MBS) overlaid with a number of small cells, inter-connected via a backhaul network deployed in a tree topology. We generalize the results available for proportional fairness under unlimited backhaul capacities to a more general objective of α-fairness, and under different scenarios of backhaul limitations. If each BS could perform its own scheduling locally, it would result in a simple operation of a HetNet. This decoupling of user scheduling processes in



(a) Different backhaul capacities, $\alpha = 1$



(b) $C_{BH} = 16Mbps$, $\alpha = 2$

Fig. 4: Performance of the two realization-agnostic heuristic schemes w.r.t. the optimal scheme, $N \in [10, 30]$

different BSs is obtained naturally in a network where the backhaul links do not have capacity limitations, and in such case, each BS can use a simple local scheduler. We have shown that if the limiting factor is the backhaul links between the MBS and the SCs, then each BS can still schedule locally and independently from the other BSs but the local scheduler can take different forms based on the level of capacity limitation. We propose a very simple scheduler that performs well under backhaul limitations.

When the link between the MBS and the core network is also a limiting factor, scheduling becomes much more complex. Each BS can still perform a local scheduling as in the previous case as long as there is a master problem that allocates feasible virtual backhaul capacities to each BS. Doing so in an optimal way is complex and expensive in terms of the amount and frequency of information exchanges but we show that a relatively simple heuristic works very well.

Appendix A
Proof of Theorem 2

The following important property of $[\mathbf{P^j_{Local}}(\omega,C_j)]$ will be useful in the ensuing analysis.

Proposition 1: If $C_j > 0$, there exists a unique optimal solution to $[\mathbf{P^j_{Local}}(\omega,C_j)]$ with $\beta_{ji} > 0$ for all $i \in A_j(\omega)$.
The proof is similar to Proposition 1 in [15].

The Lagrangian function of the local problem can be defined as follows.

$$L(\boldsymbol{\beta}_j;\mu_j,\zeta_j,\boldsymbol{l}_j) = -\sum_{i\in A_j(\omega)} U_\alpha(\beta_{ji}R_{ji}(\omega)) - \sum_{i\in A_j(\omega)} l_{j,i,\omega}\beta_{ji} +$$
$$\mu_{j,\omega}\Big(\sum_{i\in A_j(\omega)} R_{ji}(\omega)\beta_{ji} - C_j\Big) + \zeta_{j,\omega}\Big(\sum_{i\in A_j(\omega)} \beta_{ji} - 1\Big)$$

where $\boldsymbol{\beta}_j$ and $\boldsymbol{l}_j$ are respectively the vectors comprising of all β_{ji} and all $l_{j,i,\omega}$ for $i \in A_j(\omega)$. The Karush-Kuhn-Tucker (KKT) conditions [26], necessary for optimality of $[\mathbf{P^j_{Local}}(\omega,C_j)]$, can be written as follows.

$$\frac{\partial L}{\partial \beta_{ji}} = 0 \implies \beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\left(\mu_{j,\omega}R_{ji}(\omega) + \zeta_{j,\omega} - l_{j,i,\omega}\right)^{\frac{1}{\alpha}}}, \quad \forall i \in A_j(\omega) \quad (24)$$

$$\zeta_{j,\omega}\Big(\sum_{i\in A_j(\omega)} \beta_{ji} - 1\Big) = 0 \quad (25)$$

$$\mu_{j,\omega}\Big(\sum_{i\in A_j(\omega)} R_{ji}(\omega)\beta_{ji} - C_j\Big) = 0 \quad (26)$$

$$l_{j,i,\omega}\beta_{ji} = 0, \quad \forall i \in A_j(\omega) \quad (27)$$
$$\mu_{j,\omega} \geq 0; \quad \zeta_{j,\omega} \geq 0; \quad \boldsymbol{l}_j \geq 0$$
$$Eq.(13); \quad Eq.(14); \quad Eq.(15);$$

(24) are the first-order necessary conditions for optimality. (25), (26) and (27) are the so-called complementary-slackness conditions. The primal problem involves maximization of a concave function over a convex set, and hence any tuple of primal and dual variables $(\{\beta_{ji}\},\mu_{j,\omega},\zeta_{j,\omega},\{l_{j,i,\omega}\})$ that satisfies all of the KKT conditions is optimal [26]. Also, from Proposition 1, we know that such a solution is unique. Moreover, since the optimal solution is known to satisfy $\beta_{ji} > 0$, we have $l_{j,i,\omega} = 0$ for all $i \in A_j(\omega)$ from (27). Using this fact on the first order condition (24), we get

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\left(\mu_{j,\omega}R_{ji}(\omega) + \zeta_{j,\omega}\right)^{\frac{1}{\alpha}}}, \quad \forall i \in A_j(\omega) \quad (28)$$

Note that the optimal dual variables obey one of the *three* conditions: $(\mu_{j,\omega} = 0, \zeta_{j,\omega} > 0)$, $(\mu_{j,\omega} > 0, \zeta_{j,\omega} = 0)$, and $(\mu_{j,\omega} > 0, \zeta_{j,\omega} > 0)$. This is because, (28) imposes $\mu_{j,\omega}R_{ji}(\omega) + \zeta_{j,\omega} \neq 0$, for $\alpha > 0$. Hence, $(\mu_{j,\omega} = 0, \zeta_{j,\omega} = 0)$ is not possible.

We will make use of the following lemmas to establish our main result.

Lemma 2: (a) If $C_j \geq C^*_{j,\alpha}(\omega)$, then $(\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega))$ is the unique optimal solution to $[\mathbf{P^j_{Local}}(\omega,C_j)]$. (b) If $C_j < C^*_{j,\alpha}(\omega)$, then $(\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega))$ is not feasible.

Proof: It is easy to verify that $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)}$ for all $i \in A_j(\omega)$, $\mu_{j,\omega} = 0$ and $\zeta_{j,\omega} = \left(\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)\right)^\alpha$ satisfy all KKT conditions if $C_j \geq C^*_{j,\alpha}(\omega)$. It is thus an optimal solution consistent with the backhaul capacity value $C_j \geq C^*_{j,\alpha}(\omega)$. Proposition 1 implies that this is in fact the only optimal solution. If $C_j < C^*_{j,\alpha}(\omega)$, substituting $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)}$ for all $i \in A_j(\omega)$ in $\sum_{i\in A_j(\omega)} \beta_{ji}R_{ji}(\omega) \leq C_j$ results in a contradiction. ■

Lemma 3: (a) If $C_j \leq c_j^*(\omega)$, then $(\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \forall i \in A_j(\omega))$ is the unique optimal solution to $[\mathbf{P^j_{Local}}(\omega, C_j)]$. (b) If $C_j > c_j^*(\omega)$, then $(\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \forall i \in A_j(\omega))$ is not feasible.

Proof: We can easily verify that $\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}$ for all $i \in A_j(\omega)$, $\mu_{j,\omega} = \left(\frac{|A_j(\omega)|}{C_j}\right)^{\alpha}$ and $\zeta_{j,\omega} = 0$ satisfy all KKT conditions if $C_j \leq c_j^*(\omega)$. It is thus an optimal solution consistent with the backhaul capacity value $C_j \leq c_j^*(\omega)$. Proposition 1 implies that this is also the only optimal solution. If $C_j > c_j^*(\omega)$, then substituting $\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}$ for all $i \in A_j(\omega)$ in $\sum_{i \in A_j(\omega)} \beta_{ji} \leq 1$ results in a contradiction. ■

Lemma 4: If $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, the optimal dual solution is obtained by solving the following equations for $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$.

$$\sum_{i \in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = C_j \tag{29}$$

$$\sum_{i \in A_j(\omega)} \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = 1 \tag{30}$$

Proof: We will first show that the optimal dual variables have to satisfy $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$. First, we assume that there exists a dual optimal solution such that $\mu_{j,\omega} = 0$. $\mu_{j,\omega} = 0$ implies $\zeta_{j,\omega} > 0$, and hence

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\zeta_{j,\omega}^{\frac{1}{\alpha}}} \text{ and } \sum_{i \in A_j(\omega)} \beta_{ji} = 1$$
$$\implies \beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega)$$

We know from Lemma 2(b) that this is an infeasible solution since $C_j < C_{j,\alpha}^*(\omega)$. Thus, we require $\mu_{j,\omega} > 0$.

Similarly, we assume that there exists a dual optimal solution such that $\zeta_{j,\omega} = 0$. $\zeta_{j,\omega} = 0$ implies $\mu_{j,\omega} > 0$, and hence

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega))^{\frac{1}{\alpha}}} \text{ and } \sum_{i \in A_j(\omega)} \beta_{ji} R_{ji}(\omega) = C_j$$
$$\implies \beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \forall i \in A_j(\omega)$$

We know from Lemma 3(b) that this is an infeasible solution since $C_j > c_j^*(\omega)$. Thus, we require $\zeta_{j,\omega} > 0$.

Thus, the optimal solution has to satisfy $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$. In such case, (25) and (26) require that the primal constraints (13) and (14) are satisfied with equality. Substituting the value of β_{ji} from (28) in these equalities, we get the required equations. ■

Proofs for Lemma 2, 3, and 4 complete the proof for Theorem 2.

APPENDIX B
PROOF OF THEOREM 1

Note that the optimal schedules for Scenario 0 have to be equal to the solutions for sufficiently large values of C_j. So, the proof of Lemma 2 contains the proof for Theorem 1.

APPENDIX C
PROOF OF LEMMA 1

The results for $c_j^*(\omega) \leq C_j$ and $C_j \geq C_{j,\alpha}^*(\omega)$ are immediate from the closed-form solutions of $f_{j,\omega}(C_j)$ from Theorem 2.

For $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, we know that an optimal dual variable $\mu_{j,\omega}^*(C_j)$ is a subgradient of $f_{j,\omega}(C_j)$ at C_j. We need to show that this is unique and is the only subgradient, or alternatively we need to show that $f_{j,\omega}(C_j)$ is differentiable.

The differentiability of $f_{j,\omega}(C_j)$ can be shown by noting that the local problem has a unique optimal dual solution $\mu_{j,\omega}^*(C_j)$ for $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$. Applying this uniqueness in Corollary 5(ii) of [27] proves differentiability.

APPENDIX D
PROOF OF THEOREM 3

We first establish the following proposition which allows us to compute the primal variables $\{\tilde{C}_j\}_{j \in \{0\} \cup \mathcal{P}}$ that maximize the lagrangian function for a given dual variable Λ.

Proposition 2: $\tilde{C}_{j,\omega,C_j}^D(\Lambda) = \min\{f_{j,\omega}'^{(-1)}(\Lambda), C_j\}, \quad \forall j \in \{0\} \cup \mathcal{P}$ give the values of virtual capacities $\{\tilde{C}_j\}$ that maximize the Lagrangian function $L(\tilde{\boldsymbol{C}}; \Lambda)$ for a given Λ where $f_{j,\omega}'^{(-1)}(\Lambda)$ is defined in (23), with $\mu_{j,\omega}^{*(-1)}(\Lambda)$ representing the inverse mapping of $\mu_{j,\omega}^*(C_j)$ in the interval of $(0, \left(\frac{|A_j(\omega)|}{c_j^*(\omega)}\right)^{\alpha})$.

Proof: Case 1: $C_j \geq C_{j,\alpha}^*(\omega)$ We first prove the proposition for the case of large $\{C_j\}$ (specifically, $C_j \geq C_{j,\alpha}^*(\omega)$ for all j). In this case, $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j) = f_{j,\omega}(\tilde{C}_j)$. The Karush-Kuhn-Tucker (KKT) first-order conditions ($\frac{\partial L}{\partial \tilde{C}_j} = 0$) give us the following.

$$f_{j,\omega}'(\tilde{C}_j) = \Lambda \quad \forall j \in \{0\} \cup \mathcal{P} \tag{31}$$

Thus, for all $\Lambda > 0$, we require that a primal variable $\tilde{C}_j$ has to be less than or equal to $C_{j,\alpha}^*(\omega)$ (or, otherwise $f_{j,\omega}'(\tilde{C}_j)$ would be 0, which means $\Lambda = 0$). Together with this, the strictly decreasing nature of $f_{j,\omega}'(\tilde{C}_j)$ for $0 < \tilde{C}_j \leq C_{j,\alpha}^*(\omega)$ allows us to compute an inverse function of $f_{j,\omega}'(C_j)$, defined as $f_{j,\omega}'^{(-1)}(\Lambda)$, for all $\Lambda > 0$ and that, by definition, it should satisfy (31). Finding the exact description of this inverse function is not difficult, as outlined below.

The inverse function of $f_{j,\omega}'(C_j)$ with an image in $(0, c_j^*(\omega)]$ has a domain of $\Lambda \in [\left(\frac{|A_j(\omega)|}{c_j^*(\omega)}\right)^{\alpha}, \infty)$, whose expression, shown in (23), is immediate from (19). This inverse function with an image in $(c_j^*(\omega), C_{j,\alpha}^*(\omega)]$ has a domain of $\Lambda \in (0, \left(\frac{|A_j(\omega)|}{c_j^*(\omega)}\right)^{\alpha})$, and is given by the inverse of dual variable $\mu_{j,\omega}^*(\tilde{C}_j)$, since $\frac{\partial f_{j,\omega}(\tilde{C}_j)}{\partial \tilde{C}_j} = \mu_{j,\omega}^*(\tilde{C}_j)$.

For $\Lambda = 0$, $f_{j,\omega}'(\tilde{C}_j) = \Lambda$ does not have a unique solution as $f_{j,\omega}'(\tilde{C}_j) = 0$ is true for all $\tilde{C}_j \geq C_{j,\alpha}^*(\omega)$. Choosing $\tilde{C}_j = C_{j,\alpha}^*(\omega)$ as the unique map of the inverse function for $\Lambda = 0$ thus does not affect optimality.

Case 2: $C_j < C_{j,\alpha}^*(\omega)$ For $C_j < C_{j,\alpha}^*(\omega)$, the additional requirement of the inverse mapping is that the value of primal variables as a function of Λ have to be feasible. A bounded version of the inverse mapping, with an upper-bound

of C_j would satisfy the primal feasibility constraints, which is exactly what $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ guarantees. ■

Since $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ is a non-increasing function of Λ in $[0,\infty)$, and since $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ is non-decreasing in $\tilde{C}_j$, $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ can be solved my taking the smallest value of Λ so that the *MBS backhaul constraint* is satisfied. This is exactly what Theorem 3 states.

REFERENCES

[1] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 10 –21, June 2011.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 82–89, February 2014.

[3] Huawei, "The second phase of LTE-Advanced," http://www.huawei.com/ilink/en/download/HW_259010.

[4] R. Madan *et al.*, "Cell association and interference coordination in heterogeneous LTE-a cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479 –1489, December 2010.

[5] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, 2013.

[6] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, April 2006, pp. 1 –12.

[7] M. Coldery, U. Engstrom, K. W. Helmersson, M. Hashemi, L. Manholm, and P. Wallentin, "Wireless backhaul in future heterogeneous networks," *Ericsson Review*, vol. 91, Nov 2014, http://www.ericsson.com/ae/res/thecompany/docs/publications/ericsson_review/2014/er-wireless-backhaul-hn.pdf.

[8] Patrick Donegan, "Small Cell Backhaul: What, Why and How? (white paper)," [Online], Website, July 2012, http://www.tellabs.com/resources/papers/tlab_smallcellbackhaul_wp.pdf.

[9] Ericsson, "It all comes back to backhaul," [Online], Website, August 2014, http://www.ericsson.com/res/docs/whitepapers/WP-Heterogeneous-Networks-Backhaul.pdf.

[10] A. Ting, D. Chieng, K. H. Kwong, I. Andonovic, and K. Wong, "Dynamic backhaul sensitive network selection scheme in LTE-WiFi wireless hetnet," in *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 3061–3065.

[11] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 74–80, February 2014.

[12] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *Access, IEEE*, vol. 1, pp. 335–349, 2013.

[13] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," *Communications Magazine, IEEE*, vol. 50, no. 2, pp. 148–155, February 2012.

[14] China Mobile, "C-RAN: the road towards green RAN," [Online], Website, Oct 2011, http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf.

[15] J. Ghimire and C. Rosenberg, "Impact of limited backhaul capacity on user scheduling in heterogeneous networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2014, pp. 2480 – 2485.

[16] V. Jungnickel, K. Manolakis, S. Jaeckel, M. Lossow, P. Farkas, M. Schlosser, and V. Braun, "Backhaul requirements for inter-site cooperation in heterogeneous lte-advanced networks," in *2013 IEEE International Conference on Communications Workshops (ICC)*, June 2013, pp. 905–910.

[17] O. Somekh, O. Simeone, A. Sanderovich, B. Zaidel, and S. Shamai, "On the impact of limited-capacity backhaul and inter-users links in cooperative multicell networks," in *42nd Annual Conference on Information Sciences and Systems, 2008. CISS 2008*, March 2008, pp. 776–780.

[18] H. Xu and P. Ren, "Joint user scheduling and power control for cell-edge performance improvement in backhaul-constrained network mimo," in *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 1342–1346.

[19] W. Yu, T. Kwon, and C. Shin, "Joint scheduling and dynamic power spectrum optimization for wireless multicell networks," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2010, pp. 1–6.

[20] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, 2013.

[21] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking (ToN)*, vol. 8, no. 5, pp. 556–567, 2000.

[22] R. Srikant, *The mathematics of Internet congestion control*. Springer, 2004.

[23] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[24] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[25] 3GPP-TSG-RAN-WG1, "Evolved universal terrestrial radio access (EUTRA)," *3GPP, Tech. Rep. TR 36.814*, 2010.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Pr, 2004.

[27] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

Jagadish Ghimire received his B.E. degree in Electronics and Communications Engineering from Tribhuvan University (TU), Nepal, in 2007. He received the M.Sc. degree in Communication Networks and Services from Telecom SudParis, France, in 2009, and his M.E. degree in Information and Communication Technologies from Asian Institute of Technology (AIT), Thailand, in 2009. He is currently a PhD candidate in the Department of Electrical and Computer Engineering at the University of Waterloo, Waterloo, Canada. His current research activities focus on various aspects of radio resource management in heterogeneous cellular networks.

Catherine Rosenberg was educated in France (Ecole Nationale Supérieure des Télécommunications de Bretagne, Diplôme d'Ingénieur in EE in 1983 and University of Paris, Orsay, Doctorat en Sciences in CS in 1986) and in the USA (UCLA, MS in CS in 1984), Dr. Rosenberg has worked in several countries including USA, UK, Canada, France and India. In particular, she worked for Nortel Networks in the UK, AT&T Bell Laboratories in the USA, Alcatel in France and taught at Purdue University (USA) and Ecole Polytechnique of Montreal (Canada). Since 2004, Dr. Rosenberg is a faculty member at the University of Waterloo where she now holds a Tier 1 Canada Research Chair in the Future Internet. Her research interests are broadly in networking with currently an emphasis in wireless networking and in smart energy systems. She has authored over 150 papers and has been awarded eight patents in the USA. She is a Fellow of the IEEE and of the Canadian Academy of Engineering.. More information can be found at http://ece.uwaterloo.ca/∼cath/.