Caballero, P., et al. Multi-tenant radio access network slicing: statistical multiplexing of spatial loads, in *IEEE/ACM Transactions on Networking, 25(5), Oct. 2017, pp. 3044-3058*

DOI: https://doi.org/10.1109/TNET.2017.2720668

Multi-Tenant Radio Access Network Slicing: Statistical Multiplexing of Spatial Loads

Pablo Caballero, Albert Banchs, *Senior Member, IEEE,* Gustavo de Veciana, *Fellow, IEEE,* and Xavier Costa-Pérez, *Member, IEEE*

***Abstract*—This paper addresses the slicing of Radio Access Network (RAN) resources by multiple tenants, e.g., virtual wireless operators and service providers. We consider a criterion for dynamic resource allocation amongst tenants, based on a weighted proportionally fair objective, which achieves desirable fairness/protection across the network slices of the different tenants and their associated users. Several key properties are established, including: the Pareto-optimality of user association to base stations, the fair allocation of base stations' resources, and the gains resulting from dynamic resource sharing across slices, both in terms of utility gains and capacity savings. We then address algorithmic and practical challenges in realizing the proposed criterion. We show that the objective is NP-hard, making an exact solution impractical, and design a distributed semi-online algorithm which meets performance guarantees in equilibrium and can be shown to quickly converge to a region around the equilibrium point. Building on this algorithm, we devise a practical approach with limited computational, information, and handoff overheads. We use detailed simulations to show that our approach is indeed near-optimal and provides substantial gains both to tenants (in terms of capacity savings) and end-users (in terms of improved performance).**

***Index Terms*—Wireless Networks, Multi-tenant Networks, RAN-Sharing, Network slicing, Resource Allocation.**

I. Introduction

Driven by the capacity requirements forecasted for future mobile networks as well as the decreasing margins obtained by operators, infrastructure sharing has established itself as a key business model for mobile operators to reduce the deployment and operational costs of their networks (e.g., [1] reports a 280% increase in deals within the last 5 years). While passive and active sharing solutions, ranging from exclusive allocation of resources to roaming agreements, are used and have been standardized, these sharing approaches are based on fixed contractual agreements with Mobile Virtual Network Operators (MVNO) over long time periods (typically on a monthly/yearly basis). In this paper, we focus on a structured dynamic slicing approach which enables a much more efficient sharing of network resources, as envisioned by the 3GPP Network Sharing Enhancements for future mobile networks which the authors contributed to [2]. Following [3], our approach divides the infrastructure into *network slices*, assigning a different slice to each operator, and implements the sharing of network resources among operators by dynamically allocating resources to slices. Such a novel network slicing approach is expected to result in new business models and revenue sources for infrastructure providers (see, e.g., [3]). Indeed, this approach supports not only classical players (mobile operators) but also new ones such as Over-The-Top (OTT) service providers that may buy a *slice* of the network to ensure satisfactory service to their users (e.g., Amazon Kindle's support for downloading content or a pay TV channel including a premium subscription). In the literature, the term *tenants* is often used to refer to the different types of players, and *multi-tenancy* refers to approaches enabling dynamic network slicing and resource sharing for multiple tenants. For simplicity, hereafter we use the term operator in a broad sense to refer to classical (virtual) operators as well as the new players enabled by this approach.

In designing a practical solution for dynamic resource sharing among slices we face multiple challenges. To start with, we need a *sharing criterion* that not only allocates resources to operators (and their corresponding slices) fairly, but also, shares the resources of each operator fairly among its users. Furthermore, the criterion should allow for flexible sharing "levels" to meet operators' heterogeneous requirements; for practical purposes, these levels should be coarse-grained, rather than based on instantaneous resource needs. When allocating resources to an operator, one should take into account the numbers and locations of active users on the network – indeed some locations may see higher demand and (consequently) the associated resources might be scarce.

Beyond the criterion itself, designing an algorithm to implement it, while realizing timely adaptation to network changes, is also very challenging. Given the amount of information involved (including the channel quality of each user) and its dynamic nature, the algorithm should be as *distributed* as possible. Also, since the algorithm may be triggered frequently (whenever a user joins, leaves or changes its location), it should be *computationally efficient*. When adapting to network changes, the algorithm should control the number of handoffs triggered, as those may represent a high *overhead*.

Manuscript received May 8, 2016; revised December 23, 2016 and May 15 2017; accepted June 12, 2017; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor K. Psounis. Date of publication xxxx xx, xxxx; date of current version xxxx xx, xxxx. The work of P. Caballero and G. de Veciana was supported by NSF Award CNS-1343383, the work of A. Banchs and X. Costa-Pérez by the H2020-ICT-2014-2 5G NORMA (Grant Agreement No. 671584), and the work of A. Banchs was also partially supported by the Spanish project DRONEXT (Grant Agreement TEC2014-58964-C2-1-R).

P. Caballero and G. de Veciana are with the University of Texas at Austin.

A. Banchs is with the University Carlos III of Madrid and with the Institute IMDEA Networks, Spain.

X. Costa-Pérez is with NEC Europe Ltd., Germany.

Digital Object Identifier xx.xxxx/TNET.xxxx.xxxxxx

Key contributions: This paper proposes a criterion for slicing the network infrastructure amongst operators and an algorithm to allocate resources accordingly. The key contributions are as follows. In Section II, we introduce a criterion for dynamic resource sharing among operators; while the criterion has been proposed before, we provide a characterization supporting its use in a multi-tenant network setting. These properties are developed in Section II-C, providing insights on the optimality and fairness of the resulting allocations, and the benefits are studied in Section II-D, by characterizing the capacity savings by means of a closed-formula. We show that the criterion not only improves overall network utility but also that of each individual operator, thus guaranteeing that operators are not harmed by the sharing of resources amongst slices. In Section III-A, we show the criterion corresponds to an NP-hard problem, motivating the need to devise an efficient approximation algorithm which is introduced in Section III-B. The proposed algorithm is semi-online, distributed, incurs low computational complexity, and has been specifically designed to control overheads associated with handoffs and/or mobile user reassociations; we rely on several intermediate analytical results to drive the key design choices underlying our algorithm. Section IV provides a comprehensive performance evaluation based on detailed simulations, showing that (i) operators can save up to 80% capacity while providing the same quality to their users, and (ii) for a fixed capacity, we improve user performance in terms of file download times by up to 30%, among other results.

Related work: We next review and contrast our work to the state-of-the-art in (i) resource allocation based on proportional fairness, and (ii) resource sharing among operators.

Considerable research effort has been devoted to address the problem of fair resource allocation in networks. In wireline networks, fair resource allocation based on utility function maximization has been extensively studied following the seminal work of [4]. Building on this work, further algorithms for congestion control in multi-path environments have been proposed [5], [6]. Not unlike our work, these algorithms are distributed. However, they allow users to decide among multiple routes while we focus on a wireless setting where each user can only use one resource (her base station).

In the specific context of wireless networks, several approaches have been proposed [7]–[9] to the problem of resource allocation and user association based on weighted and unweighted proportional fairness, respectively. The unweighted case has been largely studied in the literature in different contexts (e.g., power control [10], interference avoidance [11]). The authors of [7] and [11] analyzed the complexity of the problem and proved the existence of polynomial time algorithms which provide an exact solution, and [9] designed a distributed algorithm with convergence guarantees. In contrast to the above, the resource allocation criterion proposed in this paper relies on *weighted* proportional fairness, with operator-specific weights; this is a more difficult problem as it is NP-hard [7] and the convergence of distributed greedy algorithms cannot be guaranteed [12].

Weighted proportional fair resource allocation in the context of wireless networks has also been studied from different angles. In [8], an algorithm with tight worst-case performance bounds is proposed, while [13] proposes an heuristic algorithm. In contrast to the distributed approach proposed in this paper, both algorithms are centralized and require the availability of the full network state information, which may be very challenging to gather in a timely manner. The authors of [14], [15] propose a Gibbs-sampling mechanism based on simulated annealing that converges to an optimal solution. However, the convergence of such mechanisms is known to be very slow and for this reason the authors resort to a more practical greedy solution. For the proposed greedy solution, the authors neither provide performance bounds nor analyze convergence; additionally, the overhead is not controlled, which limits their practical deployment. All the approaches mentioned above address the problem of a single-operator network, in contrast to our work which focuses on the slicing and sharing of resources among multiple operators.

Multi-operator network sharing has been studied from many different angles, including planning, economics, coverage, performance, etc. (see e.g. [16]–[18]). This paper focuses specifically on the design of algorithms for resource sharing among operators, which has been previously addressed by [19]–[23]; however, all these works differ substantially from ours in terms of scope, criterion or approach. In [19], [20], the optimization of the network utility follows a different criterion from the one in this paper, weighted proportional fairness, which (as we show) provides many desirable properties. The works of [21], [22] present a proportional fair formulation similar to ours; however, they do not provide a rationale for their choice, in contrast to the solid analytical arguments provided here. Furthermore, [21] does not address the algorithm design, while [22] uses a general non-linear solver that incurs a very high computational complexity (as confirmed by our results of Section IV-D). Finally, [23] follows a game theoretic approach where operators bid for resources, which results in a fundamentally different problem from the one addressed here.

In summary: (i) while there has been substantial research on proportional fair resource allocation, its application to multi-operator settings and the associated problems have not been studied, and (ii) in spite of the substantial work devoted to proportional fairness in general settings, there is a gap in the systematic study of distributed mechanisms for joint resource allocation and user association that build on analytical results.

II. Resource Allocation Criterion

In this section, we formulate the optimization problem that will drive (i) the association of users to base stations, and (ii) the allocation of base stations' resources to users. Hereafter, we refer to this optimization as the *multi-operator resource allocation* (MORA) criterion. We show analytically that the criterion satisfies desirable properties in terms of optimality and fairness, and develop a simple model to evaluate the potential sharing gains of our network slicing approach.

A. System model

We start by presenting our system model which was developed with LTE/LTE-A systems in mind, but is generally

applicable to cellular systems. Consider a network consisting of a set $\mathcal{B}$ of base stations (or sectors in case of sector antennas) that are shared by a set of operators $\mathcal{O}$. At any given time, we let $\mathcal{U}$ denote the set of users sharing the network and $\mathcal{U}_o,\ o \in \mathcal{O}$ the subsets of users belonging to each operator. An allocation of resources involves two sets of variables: (i) the association of users to base stations, denoted by $\mathbf{x} = (x_{ub} : u \in \mathcal{U}, b \in \mathcal{B})$, where each user u is associated with a single base station, i.e., $x_{ub} = 1$ for one of the base stations and 0 otherwise, and (ii) the allocation of the resources of each base station among its associated users, denoted by $\mathbf{f} = (f_{ub} : u \in \mathcal{U}, b \in \mathcal{B})$, where f_{ub} is the fraction of the base station b's resources which are allocated to user u.[1] Note that in our model we ignoring the discrete nature of such resources, and assume that $f_{u,b}$ can take any value in the continuous range [0,1].

We let $\tilde{c}_{ub}$ denote the average rate per resource unit seen by user u at base station b under current radio conditions,[2] and let $\mathcal{C}_b$ be the base station's total amount resources. Given that the user is allocated a fraction f_{ub} of the base station's resources, her rate is given by $f_{ub}\mathcal{C}_b\tilde{c}_{ub}$. For notational convenience, we define the achievable rate of the user as $c_{ub} := \tilde{c}_{ub}\mathcal{C}_b$, which yields the following rate allocation:

$$r_u(\mathbf{x}, \mathbf{f}) := \sum_{b \in \mathcal{B}} x_{ub} f_{ub} c_{ub}.$$

Note that the definition of c_{ub} actually represents an abstraction of the underlying physical resources, accounting for the various physical layer techniques (such as, e.g., power control or MU-MIMO) as well as the interference from different sources (including that of neighboring base stations). In line with similar analyses in the literature [19], [21], [22], [24]–[26], we shall assume that c_{ub} is fixed for each user and base station pair.

B. MORA criterion

In line with previous approaches [19], [21], [22], the underlying assumption behind our criterion is that operators share the cost of deploying and/or maintaining the infrastructure, and the resources received by each operator should be based on the level of its (financial) contribution to the shared network: if an operator contributes twice as much as another, it should roughly get twice the resources. To this end, each operator is assigned a *network share* $s_o \in [0, 1]$, to represent its level of contribution to the network. Without loss of generality, these shares are normalized so that $\sum_{o \in \mathcal{O}} s_o = 1$.

The proposed criterion allocates resources across operators dynamically, tracking changes in the numbers and locations of operators' mobile users and the associated transmission rates c_{ub}. When doing this, we need to make sure that (i) network resources are fairly shared among the various operators according to their share, and (ii) at the same time, the resources allocated to a given operator are fairly shared among the users of that operator. To achieve this, we follow an approach akin to that in [27][3]: we maximize the overall network utility resulting from aggregating operator utilities, where the utility of an operator is in turn the aggregation of its users' utilities. To this end, we define the overall network utility as the sum of operators' utilities weighted by the shares,

$$W(\mathbf{x}, \mathbf{f}) = \sum_{o \in \mathcal{O}} s_o\, U_o(\mathbf{x}, \mathbf{f}),$$

and the *operator utility* as the sum utility of the operator's users normalized by the number of users (where a user's utility is logarithmic in its rate),

$$U_o(\mathbf{x}, \mathbf{f}) = \frac{1}{|\mathcal{U}_o|} \sum_{u \in \mathcal{U}_o} \log(r_u(\mathbf{x}, \mathbf{f})),$$

By weighting the operator utilities with the shares, we give higher priority to operators with larger shares, and by normalizing with the number of users, we avoid that operators with more users are better off. For instance, with this choice, under uniformly loaded base stations an operator with twice the share of another one will get twice as many resources, independent of the number of users of each. Combining the above equations, one can rewrite the network utility as follows:

$$W(\mathbf{x}, \mathbf{f}) = \sum_{o \in \mathcal{O}} \sum_{u \in \mathcal{U}_o} w_u \log(r_u(\mathbf{x}, \mathbf{f})), \tag{1}$$

where the user weights w_u are defined as the operator network share divided by the current number of users of the operator, i.e., $w_u = s_o/|\mathcal{U}_o|$ (in simple terms, the network share of an operator is divided equally amongst its current users).[4]

With the above, we can now formulate the MORA optimization problem as follows. Such optimization corresponds to the weighted *proportional fair* criterion (see e.g. [4]) extended to a multi-operator setting that considers utilities of the operators, rather than the ones of the individual users:[5]

$$\max_{\mathbf{x}, \mathbf{f}} \ W(\mathbf{x}, \mathbf{f}), \tag{2a}$$

subject to:

$$r_u(\mathbf{x}, \mathbf{f}) = \sum_{b \in \mathcal{B}} x_{ub} f_{ub} c_{ub},\ \forall\ u \tag{2b}$$

$$\sum_{b \in \mathcal{B}} x_{ub} = 1 \text{ and } x_{ub} \in \{0, 1\},\ \forall\ b, u \tag{2c}$$

$$\sum_{u \in \mathcal{U}} f_{ub} x_{ub} \leq 1 \text{ and } f_{ub} \geq 0,\ \forall\ b, u. \tag{2d}$$

In the sequel we shall let $\mathbf{x}^{MORA}, \mathbf{f}^{MORA}$ denote a (possibly not unique) optimal solution to this optimization problem. This formulation provides the optimal resource allocation at

[1] For instance, in LTE/LTE-A f_{ub} denotes the fraction of physical Resource Blocks, in FDM the fraction of bandwidth and in TDM the fraction of time

[2] Note that such average rates depend on the choice of modulation and coding scheme(s) selected for the user, after averaging out short-term fluctuations.

[3] Reference [27] addresses a similar problem to ours in the context of users and flows, as it aims at allocating resources fairly to users while preserving fairness among the flows of each user.

[4] While our definition of network utility coincides with that for *weighted proportional fairness*, the criterion proposed here is fundamentally different: we consider resource allocation across time and vary the weights with the number of users, while *weighted proportional fairness* typically focuses on a static scenario and relies on fixed weights.

[5] Note that (2c) ensures that a user is associated with one (and only one) base station.

a given time under the current c_{ub} values (given by the selected modulation-coding schemes); in a dynamic setting, such allocations would be re-evaluated when any of the c_{ub} values change, due to changes in the (average) channel quality.

Note that, once MORA returns the user association $\mathbf{x}$ and resource allocation $\mathbf{f}$, physical layer techniques (such as MU-MIMO or power control) are employed to optimize performance, under the constraint that users are provided with rates proportional to the r_u values given by MORA.

C. Properties of MORA Resource Allocation

Next, we show that the MORA criterion satisfies some desirable properties both in the way base stations' resources are allocated to associated users, and the way users are associated with base stations.

1) Per-base station resource allocation: Let us first consider a general setting, where user associations to base stations are *fixed*, to see how MORA allocates base station resources. Let $\mathbf{x}^*$ be the fixed (not necessarily optimal) user to base station association. If we optimize the resource allocation $\mathbf{f}$ for this user association, i.e., $\max_{\mathbf{f}} W(\mathbf{x}^*, \mathbf{f})$ subject to (2b) and (2c), it can be seen from Lemma 5.1 of [8] that the resulting resource allocation is unique and given by $\mathbf{f}^M(\mathbf{x}^*) = (f^M_{ub}(\mathbf{x}^*) : u \in \mathcal{U}, b \in \mathcal{B})$, where

$$f^M_{ub}(\mathbf{x}^*) = \frac{w_u x^*_{ub}}{\sum_{v \in \mathcal{U}} w_v x^*_{vb}}. \tag{3}$$

Further if $\mathbf{x}^* = \mathbf{x}^{MORA}$, then $\mathbf{f}^M(\mathbf{x}^*) = \mathbf{f}^{MORA}$, i.e., we have MORA optimal allocation of network resources.

The above result is fairly intuitive. Users associated with a given base station are allocated resources proportionally to their weights w_u. This can be viewed as follows. The share of an operator represents the total budget of the operator. When assigning a weight $w_u = s_o/|\mathcal{U}_o|$ to users, this share is distributed among the operator's users, and hence the user's weight represents the budget of a user. As the resources allocated to a user are inversely proportional to the sum of weights at her base station, the sum of weights can be viewed as the cost of a unit of resource at the base station. Thus, operators with users associated with heavily loaded base stations will have to pay a higher cost (e.g, increase their network share or limit their overall number of users) or receive fewer resources.

The above shows that the number of active users that operators have on the network and their *spatial distribution* will impact the resources allocated under MORA. Indeed, allocations across base stations are coupled together through $|\mathcal{U}_o|$, i.e., an operator with a large number of active users will have lower weights and likely lower per-user allocations. At the same time, the resources obtained by an operator heavily depend on the *load* at base stations to which its users will be associated with.

2) User association: Next we study the MORA user associations. Building on the optimality of our formulation, we can show that the resource allocation resulting from MORA is Pareto-optimal, which means that for any alternative allocation $(\mathbf{x}', \mathbf{f}')$ for which $r_u(\mathbf{x}', \mathbf{f}') > r_u(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ for some u, we necessarily have $r_v(\mathbf{x}', \mathbf{f}') < r_v(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ for some $v \neq u$. Indeed, if this was not the case then $W(\mathbf{x}', \mathbf{f}')$ would be larger than $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$, which contradicts the fact that the optimal MORA allocation $(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ maximizes $W(\mathbf{x}, \mathbf{f})$.

Thus, Pareto optimality in this context means that if under some other user association choice, a user sees a higher throughput than that under MORA then there must be another user which sees a lower throughput allocation. Note that this need not always be the case. Consider, for instance, a network with $|\mathcal{U}|$ users, such that the largest c_{ub} of each user corresponds to a different base station. While the optimal allocation would associate each user to the base station with largest c_{ub}, a criterion based on local decisions that looks at users one by one may lead to a different association. The above result guarantees that this will not happen under MORA.

D. Gains and Savings

In the following we evaluate the benefits of MORA. To that end, we introduce a simple baseline – *static slicing* (SS), a proxy for not sharing resources at all.[6]

1) Static Slicing (SS) Baseline: Suppose each operator contracts for a *fixed* slice/fraction s_o of the network resources at each base station for its exclusive use. The operator can of course still optimize its users associations, $\mathbf{x}^o = (x_{ub} : u \in \mathcal{U}_o, b \in \mathcal{B})$, and allocation of resources $\mathbf{f}^o = (f_{ub} : u \in \mathcal{U}_o, b \in \mathcal{B})$, so as to maximize its utility. Specifically each operator $o \in \mathcal{O}$ can determine its user association and resource allocations based on:

$$\begin{aligned} \max_{\mathbf{x^o}, \mathbf{f^o}} \quad & U_o(\mathbf{x^o}, \mathbf{f^o}) \\ \text{subject to} \quad & r_u(\mathbf{x^o}, \mathbf{f^o}) = \sum_{b \in \mathcal{B}} x_{ub} f_{ub} c_{ub}, \ \forall u \in \mathcal{U}_o, \\ & \sum_{b \in \mathcal{B}} x_{ub} = 1, \quad \forall u \in \mathcal{U}_o, \\ & \sum_{u \in \mathcal{U}_o} f_{ub} x_{ub} \leq s_o, \quad \forall b \in \mathcal{B}, \\ & x_{ub} \in \{0,1\}, \ f_{ub} \geq 0, \quad \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_o. \end{aligned} \tag{4}$$

This is similar to MORA except limited to the operator o's current users $\mathcal{U}_o$ and the resource constraint corresponds only to the fixed slice s_o allocated to the operator at each base station. Although the user associations and resource allocations under static slicing are independently optimized by each operator, we shall let $\mathbf{x}^{SS}, \mathbf{f}^{SS}$ be a (possibly not unique) optimal choice across all operators under static slicing. Also paralleling our discussion of MORA, it is easy to show that if one fixes a feasible user association $\mathbf{x}^*$, (4) is convex and yields resource allocations given by

$$\mathbf{f}^S(\mathbf{x}^*) := (f^*_{ub}(\mathbf{x}^*) \ : \ \forall u \in \mathcal{U}, \forall b \in \mathcal{B}),$$

where

$$f^*_{ub}(\mathbf{x}^*) = \frac{x^*_{ub} s_o}{\sum_{v \in \mathcal{U}_o} x^*_{vb}} \mathbf{1}\{u \in \mathcal{U}_o\}, \tag{5}$$

[6]By *slicing* we refer to the way resources are shared (or sliced) among operators (while *resource allocation* refers to the allocation of resources to specific users). In contrast to the dynamic nature of MORA-based slicing, static slicing divides the infrastructure in fixed fractions.

i.e., this is again a weighted proportionally fair allocation of the operators' slice of the base station resources.

2) Operator Utility Gains and Protection: The *overall* network utility under MORA is clearly larger than that under the more constrained allocations possible under SS. This however does not guarantee that a given *operator's* utility under MORA is greater than that under SS. Below we show that for the *same* user association an operator utility under MORA exceeds that under SS, indicating that beyond the overall network utility, we have that each operator is indeed better off. This shows that MORA effectively protects operators when sharing their resources with other operators, which is very important to ensure that operators accept this criterion. Note that the result is completely general and holds for any possible scenario.[7]

Theorem 1. *For a given user association* $\mathbf{x}$*, MORA's resource allocation* $\mathbf{f}^M(\mathbf{x})$ *(see Eq. 3) achieves a higher utility than that of SS given by* $\mathbf{f}^S(\mathbf{x})$ *(see Eq. 5), i.e., for all* $o \in \mathcal{O}$

$$U_o(\mathbf{x}, \mathbf{f}^M(\mathbf{x})) \geq U_o(\mathbf{x}, \mathbf{f}^S(\mathbf{x})).$$

3) Capacity Savings: Next we consider the capacity savings resulting from operators sharing infrastructure. Specifically we compare the spectrum capacities, i.e., total amount of resource, required to achieve the same *average utility* per operator under MORA and SS. The aim is to give some intuition on the typical savings one might expect and its dependence on the network load, number of operators and their shares. For tractability we will examine a scenario where traffic is spatially homogenous and operators' network shares are proportional to their load.

We consider a network model in which there is a *fixed* total number of users $|\mathcal{U}|$ of which each operator contributes a fixed number of users proportional its network share s_o, i.e., $n_o = s_o|\mathcal{U}|$ which are assumed to be integer valued. Each operator's users are randomly (uniformly) distributed amongst the $|\mathcal{B}|$ base stations, so the number of users of operator o associated with base station b, is given by a random variable $N_{o,b}$, such that $N_{o,b} \sim \text{Binomial}(n_o, \frac{1}{|\mathcal{B}|})$. The total number of users at base station b is denoted by a random variable $N_b = \sum_{o\in\mathcal{O}} N_{o,b} \sim \text{Binomial}(|\mathcal{U}|, \frac{1}{|\mathcal{B}|})$. We also assume for simplicity that users have the same capacity $c_{ub} = c$ to the base stations with which they associate.

Note that under the above traffic model all users u have the *same* weight $w_u = \frac{s_o}{n_o} = 1/|\mathcal{U}|$. Thus expected overall network utility under MORA is given by:

$$\bar{W} = \mathbb{E}\left[\sum_{o\in\mathcal{O}}\sum_{b\in\mathcal{B}} N_{ob} w_u \log\left(\frac{c}{N_b}\right)\right] = \mathbb{E}\left[\sum_{b\in\mathcal{B}}\sum_{o\in\mathcal{O}} \frac{N_{ob}}{|\mathcal{U}|}\log\left(\frac{c}{N_b}\right)\right]$$
$$= \mathbb{E}\left[\sum_{b\in\mathcal{B}} \frac{N_b}{|\mathcal{U}|}\log\left(\frac{c}{N_b}\right)\right] = \frac{|\mathcal{B}|}{|\mathcal{U}|}\mathbb{E}\left[N_b \log\left(\frac{c}{N_b}\right)\right],$$

where the last equality follows by using the uniformity of traffic across base stations. Moreover, under our model the network utility $\bar{W}$ is the average utility across all users, which by symmetry is equal to the expected utility of a given operator o under MORA, i.e., $\bar{U}_o^{MORA} = \bar{W}$.

[7]The proofs of the theorems are provided in the Appendix.

Now applying Taylor's approximation to the function $x\log(c/x)$ at $\mathbb{E}[N_b]$ we obtain

$$N_b \log\left(\frac{c}{N_b}\right) \approx \mathbb{E}[N_b]\log\left(\frac{c}{\mathbb{E}[N_b]}\right) + \left[\log\left(\frac{c}{\mathbb{E}[N_b]}\right) - 1\right]\cdot (N_b - \mathbb{E}[N_b]) - \frac{1}{2\mathbb{E}[N_b]}(N_b - \mathbb{E}[N_b])^2,$$

which in turn gives

$$\mathbb{E}\left[N_b \log\left(\frac{c}{N_b}\right)\right] \approx \mathbb{E}[N_b]\log\left(\frac{c}{\mathbb{E}[N_b]}\right) - \frac{1}{2\mathbb{E}[N_b]}\text{Var}(N_b).$$

Since $N_b \sim \text{Binomial}(|\mathcal{U}|, \frac{1}{|\mathcal{B}|})$ we have that $\text{Var}(N_b) = \frac{|\mathcal{U}|}{|\mathcal{B}|}(1 - \frac{1}{|\mathcal{B}|}) \approx \frac{|\mathcal{U}|}{|\mathcal{B}|}$, and so

$$\bar{U}_o^{MORA} \approx \log\left(\frac{c}{\mathbb{E}[N_b]}\right) - \frac{|\mathcal{B}|}{2|\mathcal{U}|}. \tag{6}$$

Let Δ_o denote the extra capacity that operator o would require under SS to achieve the above utility. The expected utility experienced by operator o under SS is given by

$$\bar{U}_o^{SS} = \mathbb{E}\left[\sum_{b\in\mathcal{B}} \frac{N_{o,b}}{n_o}\log\left(\frac{s_o c(1+\Delta_o)}{N_{o,b}}\right)\right]$$
$$= \frac{|\mathcal{B}|}{n_o}\mathbb{E}\left[N_{o,b}\log\left(\frac{s_o c}{N_{o,b}}\right)\right] + \log(1+\Delta_o).$$

Again using a Taylor expansion this can be approximated as

$$\bar{U}_o^{SS} \approx \log\left(\frac{s_o c}{\mathbb{E}[N_{o,b}]}\right) - \frac{|\mathcal{B}|}{n_o}\frac{\text{Var}(N_{o,b})}{2\mathbb{E}[N_{o,b}]} + \log(1+\Delta_o).$$

Noting that $\text{Var}(N_{o,b}) \approx s_o\frac{|\mathcal{U}|}{|\mathcal{B}|} = \frac{n_o}{|\mathcal{B}|}$ we have that

$$\bar{U}_o^{SS} \approx \log\left(\frac{c}{\mathbb{E}[N_b]}\right) - \frac{|\mathcal{B}|}{2n_o} + \log(1+\Delta_o). \tag{7}$$

Finally equating the expected utilities, i.e., (6) and (7), we obtain the following estimate of the necessary extra capacity Δ_o required when static slicing rather than MORA is used:

$$\log(1+\Delta_o) \approx \frac{|\mathcal{B}|}{2n_o} \times (1 - s_o). \tag{8}$$

where under our traffic load model $n_o = s_o|\mathcal{U}|$.

This result gives a clear intuition on the possible savings resulting from sharing the infrastructure with MORA dynamic slicing. In particular, the savings increase exponentially in the product of two terms. The first is inversely proportional to the average number of users operator o has per base station, i.e., $n_o/|\mathcal{B}|$; indeed, if the operator has a large number of users, its multiplexing gain is already high without sharing the infrastructure, and hence there is little gain from sharing. The second term is large when the operator has a small network share: if its share is high, the operator is using most of the network resources and there is little sharing.

In summary, capacity savings will be highest when infrastructure is shared by a large number of operators each with a small number of users per base station. With current trends towards small cells, the number of users per base station is expected to be small, suggesting that infrastructure sharing may be particularly beneficial.

III. Approximation Algorithm for MORA

The analysis in previous section and simulations to be presented in the sequel suggest that MORA resource allocation across operators not only has desirable characteristics but will make efficient use of resources while protecting operators from one another. Unfortunately, as we show below, the complexity and information overheads associated with doing so for are already high for a static system, and excessive when operators' mobile users and associated channels are subject to constant change. In this section we further discuss the state-of-the-art algorithms to tackle MORA, and then propose an approximation algorithm based on a sequence of theoretical results and insights that support the design.

A. Complexity and State-of-the-Art Algorithms

The optimization problem underlying MORA is a *non-linear integer programming problem*, which can be shown to be NP-hard and hence there is no polynomial time algorithm unless $P = NP$.

Theorem 2. *The MORA problem is NP-hard.*

There have been a number of works in the literature devoted to solving problems similar to MORA. In particular, [8] proposes an approximation algorithm for the single operator case with guaranteed performance bounds. However, their approach is still computationally demanding; indeed, the results in Section IV-D, show that for a network with only 100 users, the algorithm takes 20 seconds on a dual-core 2.8GHz processor. Given that this would need to be executed every time c_{ub} values change or new users enter/leave the network,this seems computationally impractical. Moreover the proposed approach is centralized, so there would be a substantial information overhead to gather the c_{ub} of each user to each potential base stations, given the amount of data and dynamic nature of mobile users.

In the multi-operator setting, [22] proposes an approach based on using a standard non-linear solver to address a problem similar to MORA. Unfortunately the approach is also very complex and centralized. Indeed, our evaluation of this proposal in Section IV-D, shows that the time required to execute this algorithm increases sharply with the number of users, making it impractical at about 50 users. Moreover, [22] does not provide any analytical performance bounds.

In summary, to make dynamic multi-operator resource sharing possible, a new radically simplified approach is required. It should have low computational complexity and be based on distributed operation requiring only local information, to allow near real-time operation.

B. Algorithm design

In the following, we devise an algorithm for MORA that can be used in practical deployments. In contrast to previous approaches, our algorithm involves a low computational complexity and relies on data that can be gathered from neighboring base stations, allowing for a distributed implementation.[8]

[8]Note that, while the algorithm implementation is distributed, the logic is centralized: i.e., we assume that the algorithm is run centrally by a single entity, without the intervention of the different operators.

Given the user dynamics, i.e., joining, moving and leaving the network, an offline algorithm that computes an optimal resource allocation for a fixed set of users is impractical. Instead, we will pursue an approach that tracks users dynamics, and occasionally adjusts resource allocations by modifying current or new users' associations. Since reassociations of current users correspond to handoffs, their number should be kept to a minimum. To design such an algorithm, we need to answer

- Do we really need to reassociate users?
- Where should users be (re)associated to?
- In which order should users be reassociated?
- How many reassociations do we need?

For each of these questions, in the following we provide some theoretical analysis that eventually leads to our proposed algorithm. In all cases, once a user association $\mathbf{x}$ is set, resources at each base station are allocated according MORA's resource allocation $\mathbf{f}^M(\mathbf{x})$.

1) Need for reassociations: Following the standard terminology of online algorithms, we say that an algorithm is *online* if, upon a user joining the network, it only decides how to associate the new user, without triggering any reassociations of existing users. We say the algorithm is *semi-online* if it can further trigger reassociations of a limited number of users. Thus our first question is whether an online algorithm would suffice. The following theorem suggests that the performance of an online algorithm can be arbitrarily bad, motivating us to consider semi-online approaches.

Theorem 3. *Consider an online algorithm that triggers no reassociations of existing users. Let* $(\mathbf{x}', \mathbf{f}')$ *denote the solution resulting from this algorithm and* $(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ *a MORA optimal solution. Then,* $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - W(\mathbf{x}', \mathbf{f}')$ *cannot be bounded.*

2) Criterion for (re)associations: Next we address the question regarding how to associate, or reassociate, users to base stations. In particular, consider a *Distributed Greedy algorithm* wherein we iteratively examine (in arbitrary order) if there is a user which could change her association to increase her rate, and if this is the case, she chooses to re-associate with the base station providing the largest rate. The following result characterizes the performance of this algorithm *if* an equilibrium is reached.

Theorem 4. *Let* $(\mathbf{x}', \mathbf{f}')$ *be an equilibrium allocation for the Distributed Greedy algorithm, and* $(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ *a MORA optimal solution, then*[9]

$$W(\mathbf{x}', \mathbf{f}') \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - \log(e).$$

There exists an instance of the problem for which it holds that $W(\mathbf{x}', \mathbf{f}') = W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - \log(2)$.

Note that the above bound of $\log(e)$ is fairly close to the $\log(2)$ bound provided by [8]. This is quite remarkable, considering that the algorithm proposed in [8] is centralized and much more complex. Furthermore, the theorem shows that the bound is rather tight, as there exists a problem instance

[9]To gain some intuition on this bound, we note that a $\log(e)$ gap is equivalent to reducing the throughput of each user by a factor of e.

that provides a gap of $\log(2)$, which is quite close to the $\log(e)$ bound.

While the above theorem bounds network utility in equilibrium, we have not established the convergence of this algorithm to an equilibrium. The convergence of this type of algorithms has received substantial attention in the literature [12], [28], [29]. Indeed, since the throughput of user u is an increasing function of $c_{ub}/\sum_{v\in\mathcal{U}} w_v x_{vb}$, the Distributed Greedy algorithm can be viewed as a congestion game in which the load at a base station is given by the sum of weights of the users at the base station, $l_b = \sum_{v\in\mathcal{U}} w_v x_{vb}$, and a user seeks to minimize $a_{ub} l_b$, where $a_{ub} = 1/c_{ub}$. This game falls in the category of a singleton weighted congestion game with player-specific multiplicative constants and linear variable cost. Based on the lack of a counter-example and the existence of polynomial-time algorithms for special cases, [28] conjectures that this type of games have an equilibrium (see Conjecture 3.7 of [28]). Based on the simulations we have run for numerous instances of the game, we further conjecture that the Distributed Greedy algorithm (which implements a best response dynamics) converges to this equilibrium.

In particular, Distributed Greedy satisfies $W(\mathbf{x}', \mathbf{f}') \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - \log(e)$, while [8] proposes an algorithm that provides a throughput larger than $r_u(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})/(2+\epsilon)$ to all users, which translates into $W(\mathbf{x}', \mathbf{f}') \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - \log(2+\epsilon)$; hence, the algorithm of [8] provides only a slightly tighter bound than Distributed Greedy.

3) Order of reassociations: While our analysis of the Distributed Greedy algorithm suggests a user should (re)associate to maximize her rate, it does not indicate in which order user reassociations should be considered to speed up convergence. To address this, we consider the *Greedy Largest Gain algorithm*, which operates as the Distributed Greedy algorithm but at each iteration updates the association of the user achieving the highest gain, i.e., the one achieving the largest r_u^{new}/r_u^{old}, where r_u^{old} is the user's current throughput and r_u^{new} is the throughput she would receive under the improved association.

The following theorem shows that the Greedy Largest Gain algorithm exhibits a desirable convergence property. In particular, one can guarantee that at each iteration the network utility increases until it reaches $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - 2\log(e)$, and from then on it never decreases below $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - (2+\max_u w_u)\log(e)$. Note that Distributed Greedy does not exhibit this kind of behavior: if we select users in an arbitrary order, the network utility may decrease at any iteration (as the increase in utility of the reassociated user may be smaller than the decrease experienced by the other users).

Theorem 5. *Let $(\mathbf{x}^i, \mathbf{f}^i)$ be the solution at the i^{th} iteration of the Greedy Largest Gain algorithm and $(\mathbf{x}^{MORA}, \mathbf{f}^{MORA})$ a MORA optimal solution. Then $W(\mathbf{x}^i, \mathbf{f}^i)$ increases at each iteration until $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - 2\log(e)$, and thereafter it never decreases below $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - (2+\max_u w_u)\log(e)$.*

4) Proposed algorithm: Greedy Local Largest Gain. Based on the above considerations we now propose our algorithm for MORA, the *Greedy Local Largest Gain* algorithm. We shall first describe how it operates at a high level, and then provide a more detailed algorithmic description. When a user joins the network, she greedily joins the base station providing the largest throughput. However, as we have seen, we may need to consider triggering user reassociations. To limit their number and associated handoffs overheads we constrain these to at most m. For the first $m-1$ reassociations, users choose the base station that provides the largest throughput, but in the m^{th} the user chooses the base station so as to maximize the network utility $W(\mathbf{x}, \mathbf{f})$. In each of these steps, we select which user to reassociate (if any) based on Greedy Largest Gain criterion, but instead of considering all users in the network, involving possibly a high overhead, we restrict the selection *locally* to users associated with only two base stations (see below).

In a dynamic and time-varying setting, the algorithm needs to consider the following cases: (i) a user joins the network, (ii) leaves, or (iii) changes her location. The algorithm for a joining user is detailed in the pseudocode of next page. The rationale is as follows. In the optimal allocation, users are somehow balanced among base stations, users' weights playing a role in this balance. When a new user joins the network, the balance is broken and the base station with which the user associates may have too many users. Hence, in the first step we reassociate one of the users of this base station. In the next step, the base station that received the reassociated user may have too many users; however, depending on the weights of the joining and reassociated users, the original base station may still have too many users as well. Hence, we consider the users from the two base stations as candidates for reassociation. We repeat this, considering users from two base stations, in the subsequent steps. Finally, in the last step, to avoid that the reassociation of a user harms the overall performance, we select the base station association that maximizes the overall network utility rather than the throughput of the reassociated user.

When a user leaves the network, the algorithm is quite similar (pseudocode omitted for space reasons). When she moves, her c_{ub} values to the neighboring base stations may change; if, as a result of these changes, at some point the user would receive a larger throughput in a new base station, we reassociate her to this base station. Then, the old base station executes the same algorithm as when a user leaves the network while the new base station executes the algorithm corresponding to a joining user.

5) Controlling the number of reassociations: The remaining question is how to set the limit on the number of reassociations m, which determines the trade-off between the performance of the algorithm and reassociation overhead. Such trade-offs have been analyzed for a similar setting in [30], which aims to distribute tasks among servers (where each task can only be associated to a restricted set of servers) in such a way that the maximum load across all servers is minimized. This problem is similar to ours, with tasks and servers corresponding to users and base stations respectively, in the particular case where all users have the same w_u and c_{ub}. Not unlike their setting, the performance in this case is optimized when base station loads are as balanced as possible (i.e., the highest load is minimized). According to the analysis

Algorithm 1: GLLG user joining.

Definitions:
$r_{v,b}$: throughput of user v if she associates to b;
r_v : current throughput of user v;
$\mathcal{U}_b$: set of users associated to b, ($u \in \mathcal{U}$ s.t. $x_{u,b} = 1$);
$\mathcal{U}_{\{c \cup p\}}$: set of users associated to c or p;
$W_{u,q}$: network utility if user u associates to q;
Input: $\mathbf{x}$
User v joins the network:
$b' = \arg\max_{b \in \mathcal{B}} r_{v,b}$;
$x_{v,b'} = 1 \leftarrow$ Associate user v with base station b';
$[u^*, p^*] = \arg\max_{(u,p) \in \mathcal{U}_{b'} \times \mathcal{B}} \frac{r_{u,p}}{r_u}$;
if $r_{u^*,p^*}/r_u > 1$ **then**
  Associate user u^* with base station $p^*, x_{u^*p^*} = 1$;
else
  stop
$c = p^*$ (current base station);
$p = b'$ (previous base station);
for $m - 1$ *times* **do**
  $[u^*, q^*] = \arg\max_{(u,q) \in \mathcal{U}_{\{c \cup p\}} \times \mathcal{B}} \frac{r_{u,q}}{r_u}$;
  if $r_{u^*,q^*}/r_u > 1$ **then**
    Associate user u^* with base station $q^*, x_{u^*q^*} = 1$;
    $c \leftarrow q^*$; $p \leftarrow$ previous base station of user u^*;
  else
    stop
$W \leftarrow$ current network utility;
$[u^*, q^*] = \arg\max_{(u,q) \in \mathcal{U}_{\{c \cup p\}} \times \mathcal{B}} \frac{W_{u,q}}{W}$;
if $W_{u^*,q^*}/W > 1$ **then**
  Associate user u^* with base station $q^*, x_{u^*q^*} = 1$;

of [30], the performance in terms of the highest load with our algorithm (which has a limit of m reassociations) over the highest load with the optimal algorithm (with no constraint m) is given by $O(e^{1-\frac{m}{ln|\mathcal{B}|}})$. This shows that algorithm's performance improves rapidly (exponentially) in m, and suggests a small m suffices to achieve near-optimal network utility.

To further explore the impact of m on network utility, we present the following simulation results (see Section IV for a description of the simulation setup). Here, $W(m)$ is the network utility achieved for a given m value, $W(\infty)$ is the utility with unconstrained overhead, $W(0)$ is the utility with no reassociations, and $G_W(m) \doteq 1 - \frac{W(m)-W(\infty)}{W(0)-W(\infty)}$ represents the normalized utility gain with m reassociations, showing how close we get to the unconstrained overhead utility. Fig. 1 depicts this gain as a function of m for different scenarios. As can be seen, utility gains increase very sharply. Furthermore, for $m = 3$ the gains are already very close to their maximum value; based on this, we set m equal to 3 (this is indeed the value used in the experiments of Section IV). With this setting, the proposed algorithm only introduces a small overhead, since our approach may trigger up to three handovers for every handover performed by a "traditional" solution [31].

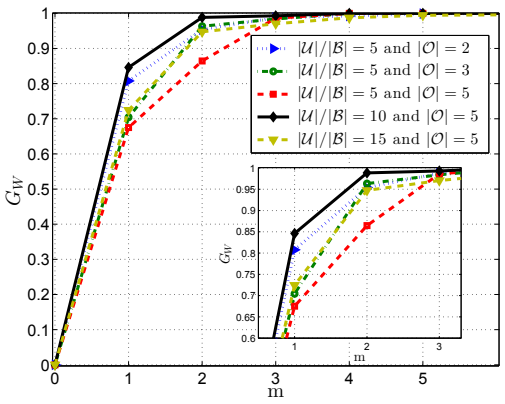

Fig. 1. Normalized utility gain as a function of m.

IV. Performance evaluation

Next, we evaluate the performance of our proposed approach. The mobile network scenario considered is based on the IMT Advanced evaluation guidelines for dense 'small cell' deployments [32]. It consists of base stations with an intersite distance of 200 meters in a hexagonal cell layout with 3 sector antennas (thus in this setting users will associate with sectors rather than the base stations we used in our algorithm description). The Signal Interference to Noise Ratio (SINR) is computed as in [25], $\text{SINR}_{ub} = P_b g_{ub} / (\sum_{k \in \mathcal{B}, k \neq b} P_k g_{uk} + \sigma^2)$, where P_b is the transmit power and g_{ub} denotes the channel gain between user u and base station b, which includes path loss, shadowing, fast fading and antenna gain. Following [32], we set $P_b = 41$ dBm, $\sigma^2 = -104$dB, path loss equal to $36.7 \log_{10}(\text{dist}) + 22.7 + 26 \log_{10}(f_c)$ for carrier frequency $f_c = 2.5$GHz, and antenna gain of 17 dBi. The shadowing factor is given by a log-normal function with a standard deviation of 8dB (as in [25]) updated every second, and fast fading follows a Rayleigh distribution dependent of the user speed and the angle of incidence (as in [33]). Achievable rates are then computed with the Shannon formula, $\text{BW} \log_2(1 + \overline{\text{SINR}}_{ub})$, for the average $\overline{\text{SINR}}_{ub}$ given by fading and shadowing [24] and a channel bandwidth of BW = 10MHz [24]. Finally, the modulation-coding scheme is selected according to the $\overline{\text{SINR}}_{ub}$ thresholds reported in [34]. Unless otherwise stated (i) users move according to the Random Waypoint Model (RWP) with speeds uniformly distributed between 0.2 and 4 m/s and pause intervals between 0 and 10 seconds, (ii) network size $|\mathcal{B}|$ is 57 sectors, (iii) all operators have the same share, and (iv) the number of users of each operator is proportional to s_o, i.e., $|\mathcal{U}_o| = |\mathcal{U}| \cdot s_o$. Confidence intervals are below 1%.

A. Utility gains

We start by evaluating the gains in terms of the overall network utility. We consider a scenario with a user density of 10 users/sector and 3 operators, and plot $W(\mathbf{x}, \mathbf{f})$ as a function of the network size $|\mathcal{B}|$. In this setting, we compare the performance of our algorithm for dynamic sharing, *Greedy Local Largest Gain* ('GLLG'), against the following approaches:

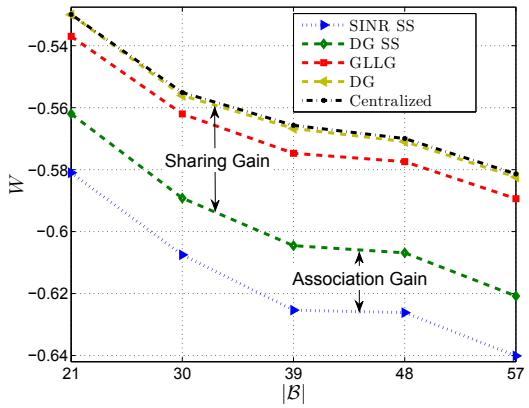

Fig. 2. Utility gains for different approaches as a function of the network size.

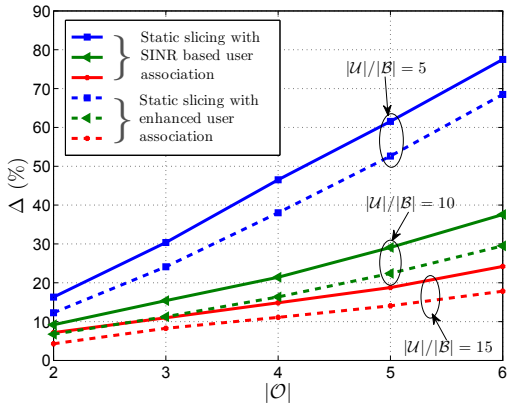

Fig. 3. Capacity savings for different scenarios as a function of the number of operators.

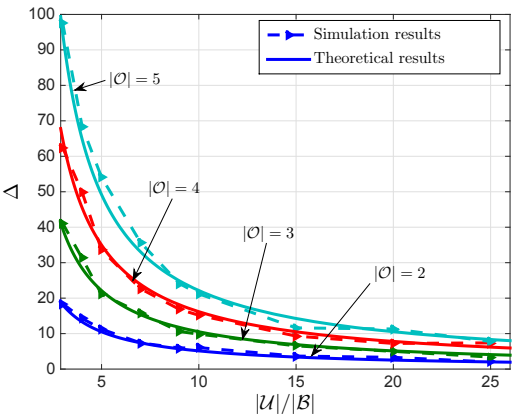

Fig. 4. Validation of the theoretical results on capacity savings.

i) *SINR-based Static Slicing* ('SINR SS'): the resources of each sector are statically divided among operators and users associate with the based station with highest SINR;
ii) *Distributed Greedy Static Slicing* ('DG SS'): resources are also sliced statically and user associations follow the Distributed Greedy algorithm discussed in Section III-B2;
iii) *Distributed Greedy* ('DG'): this is the algorithm for dynamic sharing presented in Section III-B2;
iv) *Centralized* ('Centralized'): this is the centralized algorithm proposed in [8].

The results are exhibited in Fig. 2. We draw the following conclusions: (*i*) significant gains result from both improving user association (DG SS vs. SINR SS) and sharing resources dynamically (DG vs. DG SS); (*ii*) the Distributed Greedy approach of Section III-B2 performs almost at the same level of the baseline approach of [8] (DG vs. Centralized); and (*iii*) the proposed approach performs closely to these two approaches, although it pays a small price for reducing the handoff overheads (GLLG vs. DG).

In addition to the overall network gain, it is also interesting to look at the gains of the individual operators. Theorem 1 showed that the difference in operator's utility under MORA and SS is positive as long as we have the same user association in both approaches; however, we would expect this to hold in general, i.e., even when we have different user associations. To this end, we have evaluated the difference between the operator's utility under MORA and SS over a large number of different scenarios and settings. We have observed that in all cases, MORA always provides better performance than SS to all individual operators, which confirms that MORA effectively protects all operators, ensuring gains to all of them.

B. Capacity savings

We next evaluate the benefits of our approach to operators based on the capacity savings they would achieve. Specifically, consider a network operated under our algorithm for dynamic sharing, where the capacity (i.e., total amount of resource) of each base station is given by $\mathcal{C}_{GLLG}$, and let $C_{baseline}$ be the base stations' capacity required to achieve the same network utility under two baselines: (*a*) static slicing with SINR-based user association, and (*b*) static slicing with enhanced user association (i.e., using our algorithm for user association). These two baselines allow us to study the potential gains earned due to a smarter user association and the gains achieved by dynamic resource sharing. Fig. 3 illustrates the corresponding capacity savings, computed as $\Delta = (\mathcal{C}_{baseline} - \mathcal{C}_{GLLG})/\mathcal{C}_{GLLG}$, for different numbers of operators, $|\mathcal{O}| \in \{2, \ldots, 6\}$, and three different user densities, $|\mathcal{U}|/|\mathcal{B}| = 5$ (low density), $|\mathcal{U}|/|\mathcal{B}| = 10$ (medium) and $|\mathcal{U}|/|\mathcal{B}| = 15$ (high). The results show that substantial gains can be realized, and that gains increase with the number of operators and decrease with per-sector user load. The latter is indeed rather intuitive, since under light user loads static slicing performs poorly while MORA obtains substantial benefits from statistical multiplexing.

In order to gain additional insight into the impact of the various factors, Fig. 4 displays the influence of the share of the operator (s_o) and the average load per base station sector $|\mathcal{U}|/|\mathcal{B}|$ in the percent of extra capacity required to achieve the same utility (Δ) with the static slicing with enhanced user association baseline. Results are also compared with the analytical result of Section II-D3, confirming that the theoretical analysis result holds in real conditions.

Note that in the above experiments all operators always have the same share s_o. To illustrate the behavior of MORA under heterogeneous shares, we evaluated the performance of a

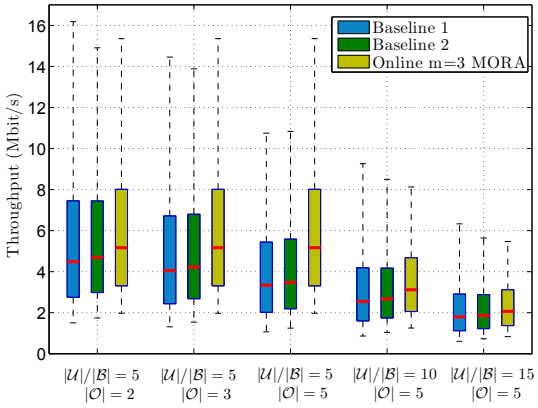

Fig. 5. Improvement on the user throughput.

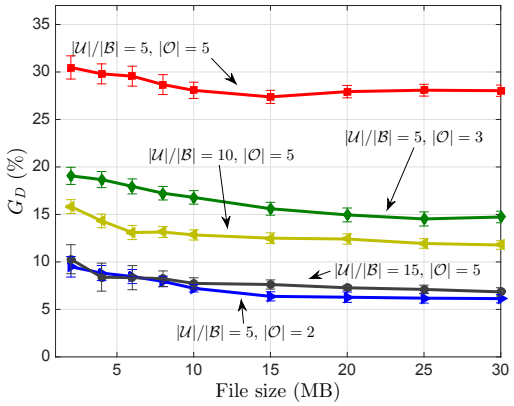

Fig. 6. Improvement on the file download time for different file sizes.

scenario with $|\mathcal{U}|/|\mathcal{B}| = 5$ and 2 operators under the following share settings: (i) $s_1 = s_2 = 1/2$ and (ii) $s_1 = 2/3$ and $s_2 = 1/3$. The gains obtained for operators 1 and 2 in the former case are $G_1 = G_2 = 11.1\%$, while in the latter case they are $G_1 = 5.3\%$ and $G_2 = 21.6\%$, respectively. Thus, this result shows that overall performance remains similar under heterogeneous shares, but gains are unevenly distributed.

C. User performance

To illustrate the gains from a user perspective, we compare the per-user throughput achieved by our approach against the two baselines: static slicing with SINR-based user association ('Baseline 1'), and static slicing with enhanced user association ('Baseline 2'). The resulting box-and-whisker plots are shown in Fig. 5 for different user densities and numbers of operators. We observe that our approach provides substantial gains both in terms of the median values as well as the various percentiles. Furthermore, as expected, gains increase with the number of operators but decrease with per-sector user load.

To complement the previous results, we compare the file download times achieved by our approach against a baseline scenario (static slicing with enhanced user association), when base stations have the same capacity in both cases and users are constantly downloading files. Let us define the file download time gain as $G_D = (D_{SS} - D_{GLLG})/D_{SS}$, where D_{SS} is the average file download time with the static slicing approach and D_{GLLG} with ours. The gains achieved are shown in Fig. 6 as a function of the file download size, for different user densities and numbers of operators. We observe the gains are substantial, and fairly independent of the file size.

D. Computational complexity

As mentioned in Section III-A, one of the key advantages of the proposed approach over the state-of-the-art is its reduced computational complexity. To quantify this, we have measured the time required to execute the following algorithms in a dual-core 2.8GHz processor: (i) our algorithm for dynamic sharing ('GLLG'); (ii) the Distributed Greedy approach of Section III-B2, which has unconstrained overhead ('DG'); (iii) the centralized algorithm of [8] ('Centralized'); and (iv) the non-linear solver used by [22] ('Non-linear Solver'). Fig. 7

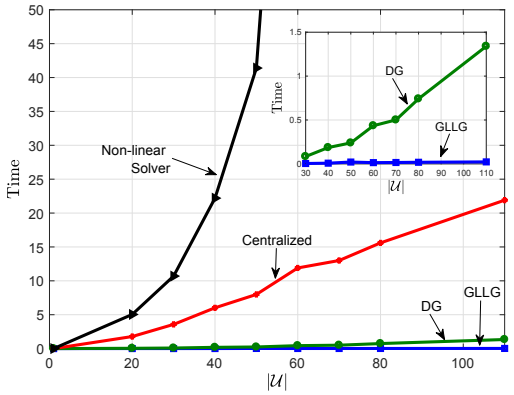

Fig. 7. Computational complexity of our approaches and state-of-the-art algorithms.

shows the resulting execution times (in seconds) as a function of the number of users for a fixed network size $|\mathcal{B}| = 57$ and $|\mathcal{O}| = 4$ operators. The results confirm that the algorithms of [22] and [8] are impractical, especially if we take into account that they have to be triggered every time the channel quality of a user changes. By contrast, the execution time of our Distributed Greedy algorithm remains very low, and it remains even lower for our GLLG approach (due to the constraint that GLLG imposes on the number of handovers).

E. Impact of non-uniform load distributions

All the results shown so far have been based on the RWP mobility model, which is known to distribute load uniformly across space. To understand the impact of non-uniform load distributions, we have evaluated the capacity savings over a baseline (static slicing with enhanced user association) under the SLAW model [35], which is a non-uniform human walk mobility model. To show different levels of non-uniformity, we have parameterized the SLAW model with five configurations of increasing non-uniformity, from $C1$ to $C5$, whose parameters {waypoints, clustering range, alpha distance, inverse self-similarity} are set as follows: $C1 = \{100, 20, 5, 0.95\}$, $C2 = \{85, 40, 4.5, 0.85\}$, $C3 = \{75, 60, 4, 0.75\}$, $C4 = \{65, 80, 3.5, 0.65\}$ and $C5 = \{50, 100, 3, 0.55\}$. The results, given in Fig. 8, show that (as expected) capacity savings decrease if loads are non-uniform, since when users concentrate

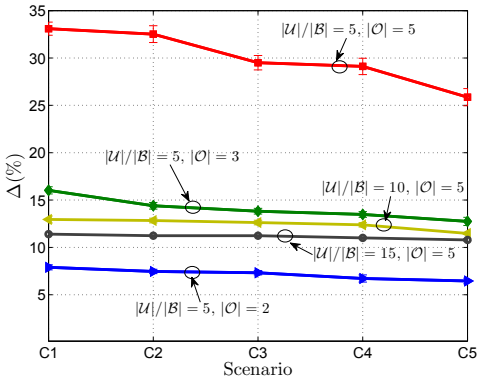

Fig. 8. Capacity savings for different levels of non-uniformity under the SLAW mobility model.

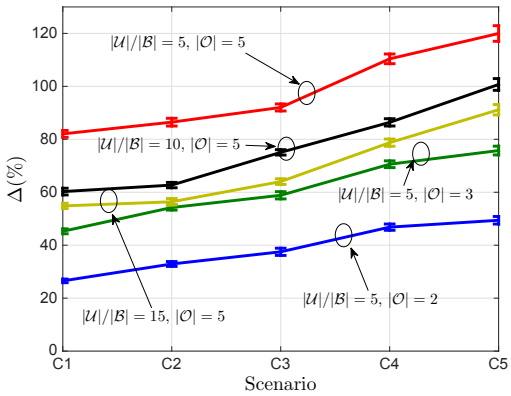

Fig. 9. Capacity savings for different levels of non-uniformity when operators follow different patterns.

around some areas the expected number of users per sector in those areas increases and thus multiplexing gains are reduced. However, the decrease is very gradual, which shows that non-uniformity has a limited impact.

The above experiment assumes that all operators follow the same mobility pattern. Alternatively, we may assume different patterns for different operators, which may be the case for instance if we consider services of different nature. To evaluate the performance under such case, we have run additional simulations in which each operator follows a different instance of the SLAW model, with different waypoints. The results, given in Figs. 9, show that in this case gains increase (rather than decrease) with non-uniformity, as each operator may have its users concentrated in different areas, thereby maximizing the benefit from resource sharing.

V. Conclusions

In this paper we have addressed the problem of multi-tenant resource slicing. While there has been substantial work towards addressing this problem, most has focused on architectural issues, leaving algorithmic aspects open to consideration. The design of algorithms for dynamic resource sharing across slices is challenging as it involves user association decisions (a difficult problem in itself) as well as multi-operator sharing policies. Our main contribution has been to show that, despite its complexity, it is possible to design practical solutions that scale to large networks and can track network load dynamics. Indeed, our analytical results provide strong evidence that the resulting allocations are near-optimal, and our simulations confirm robust benefits to operators (in terms of capacity savings) as well as to users (in terms of improved performance).

References

[1] Coleago consulting, "Mobile network sharing report," Sep. 2015. [Online]. Available: http://www.coleago.com/mobile-network-sharing-managed-services/mobile-network-sharing/
[2] 3GPP, "Telecommunication management; Network sharing; Concepts and requirements." TS 32.130, v12.0.0, Dec. 2014.
[3] NGMN Alliance, "5G White paper," Feb. 2015.
[4] F. Kelly, "Charging and rate control for elastic traffic," *European Transaction Telecommunications*, vol. 8, no. 1, pp. 33–37, Feb. 1997.
[5] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability," *Journal of the Operational Research*, vol. 49, no. 3, pp. 237–252, Mar. 1998.
[6] R. J. Gibbens and F. P. Kelly, "Resource Pricing and the Evolution of Congestion Control," *Automatica*, vol. 35, no. 12, pp. 1969–1985, 1999.
[7] T. Bu, L. Li, and R. Ramjee, "Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks," in *Proc. of IEEE INFOCOM*, Apr. 2006.
[8] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proc. of IEEE INFOCOM*, Apr. 2008.
[9] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. of IEEE INFOCOM*, Apr. 2013.
[10] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Joint resource allocation and base-station assignment for the downlink in CDMA networks," *IEEE/ACM Transactions on Networking*, vol. 14, no. 1, Feb. 2006.
[11] K. Son, S. Chong, and G. D. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3566–3576, Jul. 2009.
[12] M. Mavronicolas, I. Milchtaich, B. Monien, and K. Tiemann, "Congestion games with player-specific constants," in *Proc. of Mathematical Foundations of Computer Science*, Aug. 2007.
[13] T. Zhou, Y. Huang, W. Huang, S. Li, Y. Sun, and L. Yang, "Qos-aware user association for load balancing in heterogeneous cellular networks," in *Proc. of IEEE VTC Fall*, Sep. 2014.
[14] I.-H. Hou and C. S. Chen, "Self-organized resource allocation in LTE systems with weighted proportional fairness," in *Proc. of IEEE ICC*, May 2012.
[15] I.-H. Hou and P. Gupta, "Proportionally fair distributed resource allocation in multiband wireless systems," *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 1819–1830, Dec. 2014.
[16] P. D. Francesco, F. Malandrino, T. K. Forde, and L. A. DaSilva, "A Sharing- and Competition-Aware Framework for Cellular Network Evolution Planning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 2, pp. 230–243, Jun. 2015.
[17] B. Leng, P. Mansourifard, and B. Krishnamachari, "Microeconomic Analysis of Base-station Sharing in Green Cellular Networks," in *Proc. of IEEE INFOCOM*, Apr. 2014.
[18] J. S. Panchal, R. D. Yates, and M. M. Buddhikot, "Mobile Network Resource Sharing Options: Performance Comparisons," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4470–4482, Sep. 2013.
[19] A. Gudipati, L. E. Li, and S. Katti, "RadioVisor: A Slicing Plane for Radio Access Networks," in *Proc. of HotSDN*, Aug. 2014.
[20] P. Caballero, X. Costa-Perez, K. Samdanis, and A. Banchs, "RMSC: A Cell Slicing Controller for Virtualized Multi-Tenant Mobile Networks," in *Proc. of IEEE VTC*, May 2015.
[21] I. Malanchini, S. Valentin, and O. Aydin, "Generalized resource sharing for multiple operators in cellular wireless networks," in *Proc. of IWCMC*, Aug. 2014.
[22] R. Mahindra, M. Khojastepour, H. Zhang, and S. Rangarajan, "Radio Access Network sharing in cellular networks," in *Proc. of IEEE ICNP*, Oct. 2013.
[23] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, ""Network Slicing Games: Enabling Customization in Multi-tenant Networks," in *Proc. of IEEE INFOCOM*, May 2017.
[24] V. Sciancalepore *et al.*, "Interference coordination strategies for content update dissemination in LTE-A," in *Proc. of IEEE INFOCOM*, 2014.

[25] Q. Ye *et al.*, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[26] D. Yuhuan and G. de Veciana, "Wireless networks without edges: Dynamic radio resource clustering and user scheduling," in *Proc. of IEEE INFOCOM*, Apr. 2014.
[27] A. Banchs, "User fair queuing: fair allocation of bandwidth for users," in *Proc. of IEEE INFOCOM*, Mar. 2002.
[28] C. Georgiou, T. Pavlides, and A. Philippou, "Network uncertainty in selfish routing," in *Proc. of IEEE IPDPS*, Apr. 2006.
[29] M. Gairing, B. Monien, and K. Tiemann, "Routing (un-) splittable flow in games with player-specific linear latency functions." *Lecture Notes in Computer Science*, vol. 4051, pp. 501–512, 2006.
[30] J. Westbrook, "Load Balancing for Response Time," in *Proc. of the Third Annual European Symposium on Algorithms*, Sep. 1995.
[31] "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-Configuring and Self-Optimizing Network (SON) Use Cases and Solutions," 3GPP TS 36.902 v9.3.0, March 2011.
[32] ITU-R, "Report ITU-R M.2135-1, Guidelines for evaluation of radio interface technologies for IMT-Advanced," Technical Report, 2009.
[33] H. Dhillon *et al.*, "Modeling and Analysis of K-Tier Downlink Heterogeneous Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, Apr. 2012.
[34] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," TS 36.213, v12.5.0, Rel. 12, Mar. 2015.
[35] K. Lee *et al.*, "SLAW: self-similar least-action human walk," *IEEE/ACM Transactions on Networking*, vol. 20, no. 2, pp. 515–529, Apr. 2012.

APPENDIX

Proof of Theorem 1: For a given user association $\mathbf{x}$ the utility of operator o under SS is maximized when the resource blocks of each operator at each base station are equally distributed among the operator's users. This yields

$$U_o(\mathbf{x}, \mathbf{f}^{\mathbf{S}}(\mathbf{x})) =$$
$$= \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} \frac{1}{|\mathcal{U}_o|} x_{ub} \log\left(\frac{1}{\sum_{b\in\mathcal{B}}\sum_{v\in\mathcal{U}_o} x_{vb}} \frac{s_o}{\sum_{o'\in\mathcal{O}} s_{o'}} c_{ub}\right)$$
$$= \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\frac{1}{\sum_{b\in\mathcal{B}}\sum_{v\in\mathcal{U}_o} x_{vb}} \frac{s_o}{\sum_{o'\in\mathcal{O}} s_{o'}} c_{ub}\right)$$

where the weights are $w_u = \frac{s_o}{|\mathcal{U}_o|},\ u \in \mathcal{U}_o$.

If we multiply the numerator and denominator inside the $\log()$ by w_u, and take into account that $w_u = w_v$ for $u, v \in \mathcal{U}_o$ and $\sum_{o'\in\mathcal{O}} s_{o'} = 1$, the above can be rewritten as

$$U_o(\mathbf{x}, \mathbf{f}^{S}(\mathbf{x})) = \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log(w_u c_{ub}) - \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\frac{\sum_{b\in\mathcal{B}}\sum_{v\in\mathcal{U}_o} w_v x_{vb}}{s_o}\right).$$

The utility of operator o with MORA allocation is given by

$$U_o(\mathbf{x}, \mathbf{f}^{M}(\mathbf{x})) = \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\frac{w_u c_{ub}}{\sum_{v\in\mathcal{U}} w_v x_{vb}}\right),$$

which can be rewritten as

$$U_o(\mathbf{x}, \mathbf{f}^{M}(\mathbf{x})) = \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log(w_u c_{ub}) - \frac{1}{s_o}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\sum_{v\in\mathcal{U}} w_v x_{vb}\right).$$

From the above, if we can show that

$$\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\frac{\sum_{b\in\mathcal{B}}\sum_{v\in\mathcal{U}_o} w_v x_{vb}}{s_o}\right) \geq \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_o} w_u x_{ub} \log\left(\sum_{v\in\mathcal{U}} w_v x_{vb}\right), \quad (9)$$

the theorem is proved.

To show the above, we consider the maximization of function $\sum_{b\in\mathcal{B}} y_b \log(x_b)$ over x_b subject to $\sum_{b\in\mathcal{B}} x_b = 1$. By applying Lagrange multipliers, it can be easily seen that this function is maximized for $x_b = y_b / \sum_{b'\in\mathcal{B}} y_{b'}$. Since both the left and right-hand sides of (9) conform to this constrained optimization problem, and the left-hand side of (9) corresponds to its optimal solution, the inequality of (9) follows. □

Proof of Theorem 2: The reduction is via the 3-dimensional matching problem which is known to be NP-complete. Recall that the 3-dimensional matching problem is stated as follows. Let us consider disjoint sets $C = \{c_1, \ldots, c_n\}$, $D = \{d_1, \ldots, d_n\}$ and $E = \{e_1, \ldots, e_n\}$, and a family $T = \{T_1, \ldots, T_m\}$ of triples with $|T_i \cap C| = |T_i \cap D| = |T_i \cap E| = 1$ for $i = 1, \ldots, m$, with $m \geq n$. The question is whether T contains a matching, i.e., a subfamily T' for which $|T'| = n$ and $\cup_{T_i\in T'} T_i = C \cup D \cup E$.

Our reduction is as follows. We call the triples that contain c_j *triples of type j*. Let t_j be the number of triples of type j for $j = 1, \ldots, n$. Base station i corresponds to the triples T_i for $i = 1, \ldots, m$. We create two types of users, element users and dummy users. We have $2n$ element users, $u \in \{1, \ldots, 2n\}$, corresponding to the $2n$ elements of $D \cup E$. There are $t_j - 1$ dummy users of type j for $j = 1, \ldots, n$. Note that the total number of dummy users is $m - n$, $u \in \{2n+1, \ldots, m+n\}$. Element users can connect to the base stations that correspond to a triple that contains this element, with a transmission rate of R. Dummy users of type j can connect (also with a transmission of R) to the base stations that correspond to triples of type j. Element users have a weight $w_u = 1/(2m)$ and dummy users have a weight $w_u = 1/m$. We claim that a matching exists if and only if the network utility with the MORA criterion is $W = (n/m)\log(R/2) + ((m-n)/m)\log(R)$.

The value of the objective function is bounded above by the following optimization problem:

$$\max_{\mathbf{f}} \sum_{u=1}^{2n} \frac{1}{2m} \log(f_u R) + \sum_{u=2n+1}^{m+n} \frac{1}{m} \log(f_u R),$$

subject to $\sum_{u=1}^{2n} f_u + \sum_{u=2n+1}^{m+n} f_u = m$, where f_u is the fraction of resources assigned to user u (the first term of the summation corresponds to the element users and the second term to the dummy users).

By applying the Lagrange multiplier method, it can be easily seen that the above optimization problem is solved when $f_u = 1/2$ for the element users and $f_u = 1$ for the dummy users. This gives an upper bound on W equal to $(n/m)\log(R/2) + ((m-n)/m)\log(R)$. This corresponds to a global maximum, and thus any other set of f_u values yields a smaller W.

Assume that there is a matching. For each $T_i = (c_j, d_k, e_l)$ in the matching, we associate element users d_k and e_l with base station i. For each j, this leaves $t_j - 1$ idle base stations corresponding to tripes of type j that are not in the matching. We associate the $t_j - 1$ dummy users of type j to these $t_j - 1$ base stations. This assignment has an objective function of

$W = (n/m)\log(R/2) + ((m-n)/m)\log(R)$, which is equal to the upper bound given above. In case there is no matching, it is not possible to have the $2n$ element users sharing n base stations with $f_u = 1/2$ each, and therefore we cannot achieve the distribution of f_u values that maximizes W. According to the above result, this imples that we obtain a smaller W value. Therefore, a matching exists if and only if MORA gives $W = (n/m)\log(R/2) + ((m-n)/m)\log(R)$, which proves the theorem. □

Proof of Theorem 3: We prove the theorem by means of the following example. Let us consider a scenario with $|\mathcal{B}|$ base stations in which $|\mathcal{B}|^2$ users join the network. All users have the same weight and can associate with any of the $|\mathcal{B}|$ base stations with $c_{ub} = 1$. Independently of the criterion followed to associate new users, after all users have joined there must be a base station with at least $|\mathcal{B}|$ users. Now, suppose all users but these $|\mathcal{B}|$ leave the network. For this scenario, the network utility provided by the online algorithm is $W(\mathbf{x}', \mathbf{f}') = \sum_{i=1}^{|\mathcal{U}|} \frac{1}{|\mathcal{B}|}\log(\frac{1}{|\mathcal{B}|}) = -\log(|\mathcal{B}|)$. The optimal solution is that each user associates with a different base station, which yields $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) = \log(1)$. Thus, we have $W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - W(\mathbf{x}', \mathbf{f}') = \log(1) + \log(|\mathcal{B}|)$, which grows to ∞ as $|\mathcal{B}| \to \infty$.

Proof of Theorem 4: Since in an equilibrium of the Distributed Greedy algorithm, each user is associated with the base station that maximizes r_u, the following holds for all u:

$$\sum_{b\in\mathcal{B}} x'_{ub} w_u \log\left(\frac{w_u c_{ub}}{\sum_{v\in\mathcal{U}} x'_{vb} w_v}\right) \geq \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{w_u c_{ub}}{\sum_{v\in\mathcal{U}} x'_{vb} w_v + w_u}\right), \quad (10)$$

where the base station for which $x'_{ub} = 1$ is the one with which user u is associated under Distributed Greedy, and the base station for which $x^*_{ub} = 1$ is the one with which it is associated under the optimal allocation (i.e., $\mathbf{x}^* = \mathbf{x}^{MORA}$).

At the base station for which $x^*_{ub} = 1$ we have $\sum_{v\in\mathcal{U}} x^*_{vb} w_v \geq w_u$, so the following also holds:

$$\sum_{b\in\mathcal{B}} x'_{ub} w_u \log\left(\frac{w_u c_{ub}}{\sum_{v\in\mathcal{U}} x'_{vb} w_v}\right) \geq \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{w_u c_{ub}}{\sum_{v\in\mathcal{U}} x'_{vb} w_v + \sum_{v\in\mathcal{U}} x^*_{vb} w_v}\right).$$

Let us define the load at a base station as the sum of weights of the users at the base station, $l_b = \sum_{v\in\mathcal{U}} w_v x_{vb}$. Then, the above can be rewritten as

$$\sum_{b\in\mathcal{B}} x'_{ub} w_u \log\left(\frac{w_u c_{ub}}{l'_b}\right) \geq \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{w_u c_{ub}}{l'_b + l^*_b}\right),$$

where l'_b and l^*_b are the load at base station b with the Distributed Greedy algorithm and the optimal allocation, respectively.

From the above it follows that

$$w_u \log(r_u(\mathbf{x}^*, \mathbf{f}^*)) - w_u \log(r_u(\mathbf{x}', \mathbf{f}')) \leq \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{w_u c_{ub}}{l^*_b}\right) - \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{w_u c_{ub}}{l'_b + l^*_b}\right),$$

where $\mathbf{f}^* = f^M(\mathbf{f}^*)$. The above can be expressed as

$$w_u \log(r_u(\mathbf{x}^*, \mathbf{f}^*)) - w_u \log(r_u(\mathbf{x}', \mathbf{f}')) \leq - \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{l^*_b}{l'_b + l^*_b}\right).$$

Summing the above over all users yields

$$W(\mathbf{x}^*, \mathbf{f}^*) - W(\mathbf{x}', \mathbf{f}') \leq - \sum_{u\in\mathcal{U}} \sum_{b\in\mathcal{B}} x^*_{ub} w_u \log\left(\frac{l^*_b}{l'_b + l^*_b}\right).$$

From the above,

$$W(\mathbf{x}^*, \mathbf{f}^*) - W(\mathbf{x}', \mathbf{f}') \leq - \sum_{b\in\mathcal{B}} \log\left(\frac{l^*_b}{l'_b + l^*_b}\right)^{\sum_{u\in\mathcal{U}} x^*_{ub} w_u} = - \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} x'_{ub} w_u \log\left(\frac{l^*_b / l'_b}{1 + l^*_b / l'_b}\right)^{\frac{\sum_{v\in\mathcal{U}} x^*_{vb} w_v}{\sum_{v\in\mathcal{U}} x'_{vb} w_v}} = - \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} x'_{ub} w_u \log\left(\frac{l^*_b / l'_b}{1 + l^*_b / l'_b}\right)^{l^*_b / l'_b}.$$

Given that $(x/(1+x))^x > 1/e$ for $x \geq 0$, we obtain the following bound:

$$W(\mathbf{x}^*, \mathbf{f}^*) - W(\mathbf{x}', \mathbf{f}') \leq \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} x'_{ub} w_u \log(e) = \log(e).$$

Since $\mathbf{x}^* = \mathbf{x}^{MORA}$ and $\mathbf{f}^* = \mathbf{f}^{MORA}$, this proves the first part of the theorem.

To find an instance for which the network utility difference between MORA and Distributed Greedy Algorithm is $\log(2)$, consider the following scenario. Consider a network with 2 base stations $\mathcal{B} = \{1, 2\}$ and 2 operators $\mathcal{O} = \{1, 2\}$ with equal shares, $s_1 = s_2 = 0.5$. Each operator has one user: User 1 belongs to Operator 1 and User 2 to Operator 2. Let the achievable rates be $c_{1,1} = c_{2,2} = R$ and $c_{1,2} = c_{2,1} = R/2$, i.e, user 1 sees a higher rate with base station 1 and user 2 with base station 2. Clearly, the optimal MORA solution is to associate user 1 with base station 1 and User 2 with base station 2, i.e., $x^M_{1,1} = 1$ and $x^M_{2,2} = 1$,. This leads to a network utility $W(\mathbf{x}^{\mathbf{M}}, \mathbf{f}^{\mathbf{M}}) = 0.5\log(c_{1,1}) + 0.5\log(c_{2,2}) = \log(R)$.

Distributed Greedy Algorithm only reassociates a user if this increases her rate. Let user 1 be associated with base station 2 and user 2 with base station 2. Since none of the two users can increase her rate by reassociating, they will not reassociate with the Distributed Greedy Algorithm, and hence this algorithm will result in a user association decision $\mathbf{x}'$ such that $x'_{1,2} = 1$ and $x'_{2,1} = 1$. This yields a network utility $W(\mathbf{x}', \mathbf{f}') = 0.5\log(c_{1,2}) + 0.5\log(c_{2,1}) = \log(R/2) = \log(R) - \log(2) = W(\mathbf{x}^{\mathbf{M}}, \mathbf{f}^{\mathbf{M}}) - \log(2)$, which proves the second part of the theorem. □

Proof of Theorem 5: The proof of the theorem is based on the following steps:

Step 1: we first show that while there is some user for which $r^{new}_u \geq e \cdot r^{old}_u$, $W(\mathbf{x}^i, \mathbf{f}^i)$ increases at each iteration until we converge to a region that satisfies $r^{new}_u \leq e \cdot r^{old}_u$ for all u.

Step 2: we then show that if $r^{new}_u \leq e \cdot r^{old}_u \ \forall u$, it follows that $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{x}^{MORA}) - 2\log(e)$.

Step 3: we further prove that if a subsequent iteration i

yields $r_u^{new} \geq e \cdot r_u^{old}$ for some user u, then it must be that $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{x}^{MORA}) - (2 + \max_u w_u)\log(e)$.

Step 4: finally, we prove that after an iteration such as the above, in the subsequent iterations $W(\mathbf{x}^i, \mathbf{f}^i)$ increases, until we converge once again to a region where $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$.

We next prove each of the above steps.

Step 1: *While there is some user for which $r_u^{new} \geq e \cdot r_u^{old}$, $W(\mathbf{x}^i, \mathbf{f}^i)$ increases at each iteration until we converge to a region that satisfies $r_u^{new} \leq e \cdot r_u^{old}$ for all u.*

To prove the above, we consider a variation of the Greedy Largest Gain in which a user only moves to a new location if $r_u^{new} \geq e \cdot r_u^{old}$, and show that this algorithm is guaranteed to converge. To show this, we prove that the network utility function $W(\mathbf{x}, \mathbf{f})$ is a generalized ordinal potential for the algorithm variation. Consider the i^{th} iteration in the algorithm corresponding to a reassociation of user u, and let $(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})$ denote the configuration before this iteration and $(\mathbf{x}^i, \mathbf{f}^i)$ the configuration after the iteration. By construction of the algorithm, the following is satisfied:

$$r_u(\mathbf{x}^i, \mathbf{f}^i) \geq e \cdot r_u(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}).$$

Let b be the new base station user u associates with, and a her previous base station. Then,

$$\begin{aligned}
&W(\mathbf{x}^i, \mathbf{f}^i) - W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}) = \\
&\sum_{v \in \mathcal{U}} x_{va}^i w_v \log\left(\frac{\sum_{y \in \mathcal{U}} x_{ya}^i w_y + w_u}{\sum_{y \in \mathcal{U}} x_{ya}^i w_y}\right) + \\
&\sum_{v \in \mathcal{U}\setminus\{u\}} x_{vb}^i w_v \log\left(\frac{\sum_{y \in \mathcal{U}\setminus\{u\}} x_{yb}^i w_y}{\sum_{y \in \mathcal{U}\setminus\{u\}} x_{yb}^i w_y + w_u}\right) + \\
&w_u \log(r_u(\mathbf{x}^i, \mathbf{f}^i)) - w_u \log(r_u(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})) = \\
&l_a^i \log\left(\frac{l_a^i + w_u}{l_a^i}\right) + l_b^{i-1} \log\left(\frac{l_b^{i-1}}{l_b^{i-1} + w_u}\right) + \\
&w_u \log(r_u(\mathbf{x}^i, \mathbf{f}^i)) - w_u \log(r_u(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})).
\end{aligned}$$

Since $l_a^i \log\left(\frac{l_a^i + w_u}{l_a^i}\right) \geq 0$, we have

$$\begin{aligned}
&W(\mathbf{x}^i, \mathbf{f}^i) - W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}) \geq \\
&w_u \log\left(\frac{l_b^{i-1}/w_u}{1 + l_b^{i-1}/w_u}\right)^{\frac{l_b^{i-1}}{w_u}} + w_u \log\left(\frac{r_u(\mathbf{x}^i, \mathbf{f}^i)}{r_u(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})}\right) \\
&> w_u \log(1/e) + w_u \log(e) = 0, \qquad (11)
\end{aligned}$$

so that $W(\mathbf{x}, \mathbf{f})$ is a generalized ordinal potential. This implies that the potential game corresponding to the algorithm variation has the finite improvement property; therefore, the algorithm variation converges in a finite number of iterations to a solution that satisfies $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$. Also, from (11) it follows that $W(\mathbf{x}^i, \mathbf{f}^i) > W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})$, i.e., the network utility increases at each iteration.

As the Greedy Largest Gain algorithm always selects the user with the largest r_u^{new}/r_u^{old}, it will select a user for which $r_u^{new} \geq e \cdot r_u^{old}$, as long as there is one that satisfies this condition, and hence will follow the same steps as the algorithm variation that we have considered above. This implies that there will be some iteration i in which the Greedy Largest Gain algorithm will reach a solution $(\mathbf{x}^i, \mathbf{f}^i)$ that satisfies $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$ and, until reaching this solution, $W(\mathbf{x}^i, \mathbf{f}^i)$ will increase at each iteration.

Step 2: *If $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$, it follows that $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{x}^{MORA}) - 2\log(e)$.*

Let $(\mathbf{x}^i, \mathbf{f}^i)$ be the solution at the i^{th} iteration which satisfies $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$. Equation (10) for this solution can be rewritten as

$$\begin{aligned}
&\sum_{b \in \mathcal{B}} x_{ub}^i w_u \log\left(\frac{w_u c_{ub}}{\sum_{v \in \mathcal{U}} x_{vb}^i w_v}\right) \geq \\
&\qquad \sum_{b \in \mathcal{B}} x_{ub}^{MORA} w_u \log\left(\frac{w_u c_{ub}}{\sum_{v \in \mathcal{U}} x_{vb}^i w_v + w_u}\right) - w_u \log(e).
\end{aligned}$$

Starting from the above equation and applying the same reasoning as in the proof of Theorem 4 yields $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - 2\log(e)$.

Step 3: *If a subsequent iteration i yields $r_u^{new} \geq e \cdot r_u^{old}$ for some user u, then it must be that $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{x}^{MORA}) - (2 + \max_u w_u)\log(e)$.*

Let us that for some iteration i of the algorithm such that it holds $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$ for the solution before this iteration, and $r_u^{new} \leq e \cdot r_u^{old}$, for some u, for the solution after the iteration. Let $(\mathbf{x}^{i-1}, \mathbf{f}^{i-1})$ be the solution before iteration i and $(\mathbf{x}^i, \mathbf{f}^i)$ the solution after the iteration. As we have seen above, for the former it holds $W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - 2\log(e)$. Let us consider that at iteration i user u moves to base station b. Then,

$$\begin{aligned}
&W(\mathbf{x}^i, \mathbf{f}^i) - W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}) \geq \\
&\qquad \sum_{v \in \mathcal{U}} x_{vb}^{i-1} w_v \log\left(\frac{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t}{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t + w_u}\right) = \\
&\qquad w_u \log\left(\frac{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t / w_u}{1 + \sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t / w_u}\right)^{\frac{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t}{w_u}} \geq \\
&\qquad - w_u log(e) \geq - \max_u w_u log(e).
\end{aligned}$$

Thus,

$$W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - (2 + \max_u w_u)\log(e).$$

Step 4: *After an iteration such as the above, in the subsequent iterations $W(\mathbf{x}^i, \mathbf{f}^i)$ increases, until we converge once again to a region where $r_u^{new} \leq e \cdot r_u^{old}\ \forall u$.*

Let us consider that before iteration i there is some u for which $r_u^{new} \geq e \cdot r_u^{old}$. Then,

$$\begin{aligned}
&W(\mathbf{x}^i, \mathbf{f}^i) - W(\mathbf{x}^{i-1}, \mathbf{f}^{i-1}) \geq w_u \log(r^{new}) - w_u \log(r^{old}) + \\
&\qquad \sum_{v \in \mathcal{U}} x_{vb}^{i-1} w_v \log\left(\frac{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t}{\sum_{t \in \mathcal{U}} x_{tb}^{i-1} w_t + w_u}\right) > \\
&\qquad w_u \log(e) - w_u \log(e) \geq 0.
\end{aligned}$$

Therefore, if at some iteration we get $r_u^{new} \geq e \cdot r_u^{old}$ for some u, then for that iteration it will hold $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - (2 + \max_u w_u)\log(e)$, and from this point on $W(\mathbf{x}^i, \mathbf{f}^i)$ is going to increase until we reach $W(\mathbf{x}^i, \mathbf{f}^i) \geq W(\mathbf{x}^{MORA}, \mathbf{f}^{MORA}) - 2\log(e)$ again. □

Pablo Caballero received the B.S. ('13) and M.S. ('15) degree in telecommunications and telematics engineering respectively from the University Carlos III of Madrid. In 2015, Pablo joined the Wireless Networking and Communications Group at the University of Texas at Austin to pursue his Ph.D under the supervision of Prof. Gustavo de Veciana and Prof. Albert Banchs. Previously, Pablo worked as Research Assistant at IMDEA Networks Institute and as Research Intern at NEC Laboratories Europe. His research interests lie in the design and performance evaluation of communication networks, game theory and algorithm analysis.

Albert Banchs (M'04-SM'12) received the M.Sc. and Ph.D. degrees from the Polytechnic University of Catalonia (UPC-BarcelonaTech) in 1997 and 2002, respectively. He is currently a Full Professor with the University Carlos III of Madrid (UC3M), and has a double affiliation as Deputy Director of the IMDEA Networks institute. Before joining UC3M, he was at ICSI Berkeley in 1997, at Telefonica I+D in 1998, and at NEC Europe Ltd. from 1998 to 2003. Prof. Banchs is Editor of IEEE Transactions on Wireless Communications and IEEE/ACM Transactions on Networking. His research interests include the performance evaluation and algorithm design in wireless and wired networks.

Gustavo de Veciana (S'88-M'94-SM'01-F'09) received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively, and joined the Department of Electrical and Computer Engineering where he is currently a Cullen Trust Professor of Engineering. His research focuses on the analysis and design of communication and computing networks; data-driven decision-making in man-machine systems, and applied probability and queueing theory. Dr. de Veciana served as editor and is currently serving as editor-at-large for the IEEE/ACM Transactions on Networking. In 2009 he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He currently serves on the board of trustees of IMDEA Networks Madrid.

Xavier Costa-Pérez (M'01) is Head of 5G Networks R&D at NEC Laboratories Europe, where he manages several projects focused on 5G mobile core, backhaul/fronthaul and access networks. His team contributes to NEC projects for products roadmap evolution, to European Commission R&D collaborative projects as well as to open-source projects and related standardization bodies, and has received several R&D Awards for successful technology transfers. Dr. Costa-Pérez has served on the Program Committees of several conferences and holds multiple patents. He received both his M.Sc. and Ph.D. degrees in Telecommunications from the Polytechnic University of Catalonia (UPC-BarcelonaTech) and was the recipient of a national award for his Ph.D. thesis.