Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results

Salvatore D'Oro ⬡, *Member, IEEE*, Francesco Restuccia ⬡, *Member, IEEE, ACM*,
Tommaso Melodia ⬡, *Fellow, IEEE, Senior Member, ACM*, and Sergio Palazzo, *Senior Member, IEEE*

Abstract—**Radio access network (RAN) slicing is an effective methodology to dynamically allocate networking resources in 5G networks. One of the main challenges of RAN slicing is that it is provably an NP-Hard problem. For this reason, we design near-optimal low-complexity distributed RAN slicing algorithms. First, we model the slicing problem as a congestion game, and demonstrate that such game admits a unique *Nash equilibrium* (NE). Then, we evaluate the *Price of Anarchy* (PoA) of the NE, i.e., the efficiency of the NE as compared with the social optimum, and demonstrate that the PoA is upper-bounded by $3/2$. Next, we propose two fully-distributed algorithms that provably converge to the unique NE without revealing privacy-sensitive parameters from the slice tenants. Moreover, we introduce an adaptive pricing mechanism of the wireless resources to improve the network owner's profit. We evaluate the performance of our algorithms through simulations and an experimental testbed deployed on the Amazon EC2 cloud, both based on a real-world dataset of base stations from the OpenCellID project. Results conclude that our algorithms converge to the NE rapidly and achieve near-optimal performance, while our pricing mechanism effectively improves the profit of the network owner.**

Index Terms—**Network slicing, 5G, congestion games, game theory, distributed algorithms.**

I. INTRODUCTION

THANKS to the ubiquitousness of modern smartphones and the rise of the Internet of Things (IoT), the number of mobile devices has seen an unprecedented expansion over the last few years. The latest report by Ericsson Mobility forecasts that the number of 5G subscriptions will exceed half a billion by the end of 2022, including more than 1.5 billion IoT devices equipped with cellular connections [1]. This massive increase in the number of connected devices will necessarily result in a staggering growth in cellular data traffic. For this reason, 5G cellular networks are expected to meet stringent requirements on ubiquitous connectivity, extremely low latency, and very high-rate data transfer [2]–[4].

What 5G systems are going to be has yet to be determined. However, there is wide consensus that 5G systems will single-handedly cater to a plethora of different services, such as automotive, mobile broadband, and tactile Internet, just to name a few. Each service will have its own networking requirements, which will ultimately require the network to be *polymorphic* and seamlessly adapt to different constraints. To address this challenging problem, the notion of *network slicing* [5]–[9] has been recently proposed, where the physical and computational resources of the network are seen as a unique "object" that can be "sliced" and "served" according to an entity's current needs. In this way, heterogeneous requirements can be served by the same infrastructure in a cost-effective manner, as different network slice instances can be orchestrated and configured according to the specific requirements of the slice tenants.

Network slicing does not just provide better network performance, but it also enables realistic and profitable business models in the mobile network ecosystem. In this paper, we consider the problem of network slicing in the context of radio access networks (RANs), where telco operators (TOs) provide physical network resources to mobile virtual network operators (MVNOs), who periodically rent slices to provide cellular network access to mobile users (MUs). While a similar model is currently being successfully applied by Amazon Web Services or Microsoft Azure in the context of cloud services, slicing RAN resources is an intrinsically different problem, since (i) spectrum is a scarce resource for which over-provisioning is not possible; (ii) the network capacity is dynamic and heavily depends on the location of the MUs, among other factors; and (iii) the agreements with MVNOs usually impose stringent requirements on the Quality of Experience (QoE) perceived by the MUs. These issues motivate our paper.

The main issue in designing RAN slicing algorithms is that the most realistic problem formulations are provably NP-hard [10]. For this reason, most of existing work has focused only on the architectural aspects of network slicing [11], [12], with a limited focus on algorithmic aspects. Only very recently have a number of centralized algorithms for RAN slicing been proposed [10], [13]. Although these algorithms achieve optimality, they do not scale with the number of MVNOs and RAN resources. This issue calls for the design and analysis of *near-optimal, lower-complexity* algorithms for RAN slicing. Another crucial aspect is that optimality is often achieved to the detriment of privacy. Specifically, it is assumed that the TO is provided with complete information regarding the MVNOs' system parameters and preferences, for example, monetary budget, number and location of served MUs, business strategies, and so on. Instead, in realistic scenarios MVNOs may be reluctant to disclose such sensitive information to the TOs.

Finally, centralized solutions do not consider the competitive and selfish behavior of modern MVNOs, who will dynamically adopt different strategies to maximize their own utility without considering the broader needs of the network as a whole. This implies that *distributed* algorithms are in general more desirable than centralized solutions, as they are more likely to be adopted by MVNOs in real-world RAN slicing scenarios.

In this paper, we investigate the challenging and novel problem of *designing privacy-preserving, low-complexity, near-optimal distributed algorithms for RAN network slicing*, where the MVNOs selfishly compete with each other to acquire slices from the TOs while minimizing their cost. First, we first mathematically formulate this problem; then, we rely on the game-theoretical framework of congestion games [14]–[16] to effectively model and analyze the competition among MVNOs. Next, we demonstrate the existence and uniqueness of the *Nash equilibrium* (NE) associated to the game, and prove that its *Price of Anarchy* (PoA), i.e., the efficiency of the NE as compared to the social optimum, is limited to $\frac{3}{2}$. We formulate two distributed algorithms that preserve the privacy of the MVNOs and provably converge to the unique NE. Moreover, we propose a pricing algorithm based on a stochastic iterative mechanism that allows the TO to optimize its profit by adapting pricing policies to the current load on the resources being sliced. Finally, we extensively evaluate the performance of our algorithms through simulations and a practical testbed implemented on the Amazon EC2 cloud [17], both based on a real-world dataset of base stations (BSs) from the OpenCellID project [18]. Results demonstrate the effectiveness of the proposed approach, as our algorithms converge to the NE rapidly while achieving near-optimal performance.

II. NETWORK SCENARIO AND PROBLEM OVERVIEW

In this paper, we consider the network scenario depicted in Figure 1, which is composed by a core network (CN) and a radio access network (RAN). The CN is connected to the Internet and is in charge of routing uplink and down-link traffic. Furthermore, the RAN is composed of multiple heterogeneous remote radio heads (RRHs). We assume that both the CN and the RAN are owned and managed by a single TO. However, our model can also be applied to the more general case of multiple infrastructure owners. Given that RRHs are geographically located at different areas of the network, the RAN can be divided into multiple *RAN clusters*, which group together RRHs that are close to each other. In the considered scenario, the TO dynamically leases the network infrastructure to multiple MVNOs, who do not own any infrastructure and build virtual RANs to provide MUs with networking capabilities [6]. Since RAN clusters are geographically isolated and autonomous, slicing policies are enforced on each RAN cluster individually.

We consider the problem of generating autonomous *virtual* RANs on top of the resources provided by the physical RAN owned by the TO. A virtual RAN is thus referred to as a *network slice*, since it represents a virtual "portion" of the physical RAN. Resources that may be assigned in each slice include available spectrum, antennas, computing resources, and so on. In this scenario, it is realistic to assume the MVNOs will need to provide the TO with some sort of payment based on the number and type of resources allocated to their slices. Thus, the slicing problem reduces to finding an allocation policy for each MVNO, to minimize the cost associated to
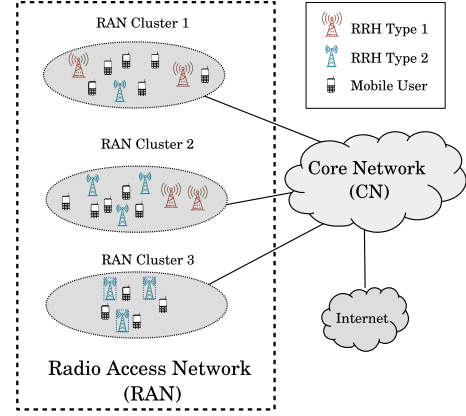

Fig. 1.   The considered 5G network scenario.

its slice while ensuring that a predetermined number of MUs can be served. Within each slice, MVNOs are free to assign resources to the MUs according to their internal policies.

This problem can be approached in two different ways. In the first approach, which we call *centralized*, the TO collects information with respect to the MUs that each MVNO would like to serve (e.g., position and number of MUs, QoS requirements, etc...), referred to as *MVNO user load*. Then, the TO computes the slices by solving an optimization problem that takes the MVNO user loads into account. Although simple in nature, this approach suffers from a number of disadvantages. First, it requires the MVNOs to disclose both business models and information related to their users to the TO, which clearly raises privacy and marketing concerns as the MVNOs are usually interested in hiding such information from external entities. Second, since centralized approaches find the global optimum by mathematical optimization [10], they require significant computational resources as the network slicing problem becomes intractable as the number of MVNOs and resources increases [13]. Thus, since the MVNO user load is dynamic in nature and business strategies might change over time, the centralized approach may fail to provide slices before their user load changes again and a different slice is thus needed. In this paper, we take a different approach and tackle the problem in a *distributed* way. Specifically, we design distributed slicing algorithms such that each MVNO computes its slice locally *without revealing relevant information, e.g., business model and position of served MUs, to the TO or other MVNOs*.

Although such distributed solution might be sub-optimal, we show that it is low-complexity (i.e., scalable) and approximates the optimum solution by a factor of $\frac{3}{2}$, a very small number. We use the framework of congestion games [14]–[16] to ensure that the MVNOs, although selfish and competing with each other, will nevertheless reach a solution (i.e., Nash equilibrium) that will satisfy each MVNO. In order to increase the TO's profit, we propose an adaptive resource pricing algorithm.

III. RAN SLICING PROBLEM FORMULATION

Let us define $\mathcal{M}$ and $\mathcal{H}$ as the sets of the MVNOs and RAN clusters, respectively. We assume that a set $\mathcal{R}_h$ of R heterogeneous RRHs are deployed for each cluster $h \in \mathcal{H}$, and that RRHs can be classified into L classes according to their performance and resource availability (e.g., macro, micro, pico and femto cells). Accordingly, the set of RRHs in RAN cluster

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

D'ORO *et al.*: LOW-COMPLEXITY DISTRIBUTED RADIO ACCESS NETWORK SLICING

3

h is $\mathcal{R}_h = (\mathcal{R}_h^{(1)}, \mathcal{R}_h^{(2)}, \ldots, \mathcal{R}_h^{(L)})$. For each RRH $r \in \mathcal{R}_h$, we also assume that the TO fixes a leasing price, here denoted as p_r.

It is reasonable to assume that statistical information on the expected number of MUs and their position in each cluster may be obtained by exploiting real-time monitoring of the RAN [19]. Given such statistics are time-varying and depend on MUs mobility patterns, we assume that MVNOs change their slicing policy on regular basis. For these reasons, we assume time is discretized into slots, where $t = 1, 2, \ldots$ represents the slot index, and the number of MUs and their position in each cluster is assumed to vary at each slot.

Let Λ_r be the maximum amount of resources of RRH r available at each slot,[1] and let N_r denote the number of MUs that can be served by RRH r without any degradation in QoE. In other words, if the number of MUs served by RRH r is lower than N_r, then all MUs connected to r enjoy high-quality communication. On the other hand, MUs would experience congestion and/or poor performance if their number is equal or greater than N_r.

The N_r parameter is specific to a given RRH and depends on Λ_r, the desired minimum QoE level, the position of MUs and their number at a given slot. As an example, it is easy to observe that communication between distant RRHs and MUs is expected to result in more collisions and errors, which necessarily increases the number of re-transmissions and the transmission power. In general, the value of N_r can be either fixed by hardware [22] or estimated by the TO by analyzing statistical information on the position, number and distribution of MUs collected by using historical data, anticipatory and prediction models [23]–[26]. The above discussion shows that *i) the value of N_r depends on a variety of heterogeneous and scenario-specific parameters, and ii) there is no unique way of characterizing such variable*. For this reason, in this paper we keep a general approach at the problem and do not make any limiting assumption on the value of N_r. Instead, we consider it as a scenario-specific parameter whose value is fixed and disclosed by the TO to the MVNOs according to a particular network configuration.

However, since N_r is expected to considerably impact the RAN slicing problem, in Section VIII we describe a methodology to obtain a worst-case estimation of the value of N_r that accounts for QoE requirements, resource availability, position and number of MUs. Our derivation is not intended to be precise and accurate, but it is only used to investigate the impact of N_r on the achievable performance of the system.

For the sake of simplicity, henceforth we will focus on a given cluster $h \in \mathcal{H}$, and we will omit both c and the slot index t. We summarize the most relevant system parameters in Table I.

A. Definition of Slicing Policy

We define as *slicing policy* a function determining how each MVNO allocates its MUs to the available RRHs in a given RAN cluster. More formally, a slicing policy for an MVNO m is an *allocation vector* $\boldsymbol{\xi}_m = (\xi_{m,r})_{r \in \mathcal{R}} \in \mathcal{S}_m$, where $\mathcal{S}_m$

TABLE I

SUMMARY OF NOTATION

Variable	Description
$\mathcal{M}, \mathcal{R}, \mathcal{H}$	Sets of MVNOs, RRHs and RAN Clusters
M, R	Number of MVNOs and RRHs
$\mathcal{R}_h$	Set of the L classes of RRHs in cluster h
N_r	Num. of MUs served by RRH $r \in \mathcal{R}$ with guaranteed QoE
p_r	Leasing price for RRH $r \in \mathcal{R}$
$t = 1, 2, \ldots$	Slicing slot index
$n_{m,h}(t)$	Expected number of MUs in cluster h for MVNOs m at slot t
$\xi_{m,r}(t)$	Number of MUs of MVNO m served through RRH r
$\Gamma_m(\boldsymbol{\xi}_m)$	Set of RRHs in $\mathcal{R}$ selected by MVNO m
$\Theta_r(\boldsymbol{\xi})$	Set of MVNOs that have selected RRH r under policy $\boldsymbol{\xi}$
$\boldsymbol{\xi}_m(t)$	Allocation vector for MVNO m
$\boldsymbol{\xi}(t)$	Allocation policy at slot t
Λ_r	Amount of resources available at RRH r
$\sigma_{m,r}(\boldsymbol{\xi})$	Available resources allocated to MVNO m under allocation policy $\boldsymbol{\xi}$
$\phi_r(\boldsymbol{\xi})$	Congestion level on RRH $r \in \mathcal{R}$ under allocation policy $\boldsymbol{\xi}$
$\pi_m^{\mathcal{P}}$	Weighing parameter for MVNO m
$c_m^C(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m})$	Congestion cost for MVNO m under allocation policy $\boldsymbol{\xi}$
$c_m^{\mathcal{P}}(\boldsymbol{\xi}_m)$	Monetary costs for MVNO m under allocation policy $\boldsymbol{\xi}$
$c_m(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m})$	Cost function for MVNO m under allocation policy $\boldsymbol{\xi}$
$\mathcal{G}$	RAN slicing congestion game
$\Phi(\boldsymbol{\xi})$	Potential function for the game $\mathcal{G}$ under allocation policy $\boldsymbol{\xi}$
$\boldsymbol{\xi}^{\text{OPT}}, \boldsymbol{\xi}^{\text{NE}}$	Optimal and Nash Equilibrium allocation policies
$C^{\text{NE}}, C^{\text{OPT}}$	Social welfares computed under the NE and optimal allocation policies
γ_n	Step-size parameter used in the learning procedure in (21)
$\Pi(t)$	Profit ot the TO at slot t

is a $(R-1)-$simplex defined as

$$\mathcal{S}_m = \left\{ (\xi_{m,r})_{r \in \mathcal{R}} : \sum_{r \in \mathcal{R}} \xi_{m,r} = n_m \text{ and } \xi_{m,r} \geq 0, \; \forall r \in \mathcal{R} \right\} \tag{1}$$

and $\xi_{m,r}$ represents the number of MUs in the cluster that the MVNO m expects to serve through RRH r.

Ultimately, the TO generates the global *allocation policy* $\boldsymbol{\xi} = (\boldsymbol{\xi}_m)_{m \in \mathcal{M}}$ by considering all the allocation vectors for each cluster. For any given allocation policy $\boldsymbol{\xi}$, the TO generates a corresponding network slice over the available RRHs. Specifically, the amount of resources at RRH $r \in \mathcal{R}$ provided to MVNO m under allocation policy $\boldsymbol{\xi}$ is:

$$\sigma_{m,r} = s(\Lambda_r, \boldsymbol{\xi}), \tag{2}$$

where Λ_r is the maximum amount of resources of RRH r, and $s(\cdot)$ is the *slicing function*. For example, a simple (yet effective), slicing function can be provided by the *proportional allocation function* [27]

$$\sigma_{m,r} = \frac{\xi_{m,r}}{\sum_{l \in \mathcal{M}} \xi_{l,r}} \Lambda_r. \tag{3}$$

Since Λ_r is a positive and finite real number, the relationship $\sum_{m \in \mathcal{M}} \sigma_{m,r} \leq \Lambda_r$ always holds for each $r \in \mathcal{R}$. For any $\boldsymbol{\xi}$, a slicing rule $\boldsymbol{\sigma}_r = (\sigma_{m,r})_{m \in \mathcal{M}}$ for each RRH $r \in \mathcal{R}$ is derived, and a *slicing policy* $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_r)_{r \in \mathcal{R}}$ is obtained. For the sake of simplicity, we also define $\Gamma_m(\boldsymbol{\xi}_m)$ as the set of RRHs in $\mathcal{R}$ selected by MVNO m in its slice, and $\Theta_r(\boldsymbol{\xi})$ as the set of MVNOs which have selected RRH r under slicing policy $\boldsymbol{\xi}$. More formally, $\Gamma_m(\boldsymbol{\xi}_m) = \{r \in \mathcal{R} : \xi_{m,r} > 0, \xi_{m,r} \in \boldsymbol{\xi}_m\}$ and $\Theta_r(\boldsymbol{\xi}) = \{m \in \mathcal{M} : \xi_{m,r} > 0, \xi_{m,r} \in \boldsymbol{\xi}\}$.

IV. RAN SLICING AS A CONGESTION GAME

When designing optimum slicing policies, MVNOs must also take into account aspects such as (i) the cost incurred when leasing RRHs; and (ii) QoS-related metrics such as

---

[1]In our model, we do not assume any specific model for Λ_r, since it models a number of factors including computational, storage and communication resources. For example, Λ_r might represent the number of OFDM symbols available in OFDM-based network architectures such as LTE [20], as well as spectrum or antennas [21].

4

IEEE/ACM TRANSACTIONS ON NETWORKING

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
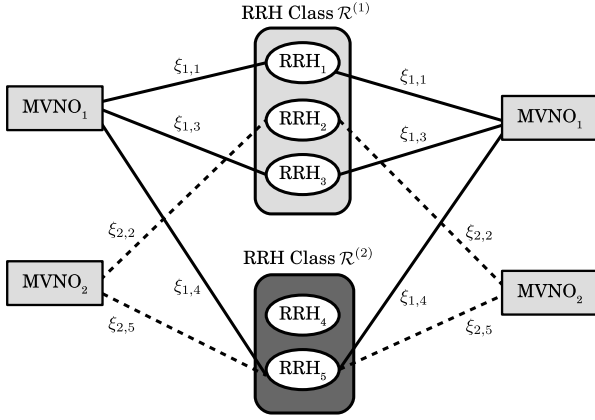

Fig. 2. Example of the considered CG on a given cluster with $|\mathcal{M}| = 2$ MVNOs and $|\mathcal{R}| = 5$ heterogeneous RRHs.

(a) the distance of those RRHs from the MUs, (b) the congestion level on each RRH, and (c) the availability of resources at each RRH. Since MVNOs are assumed as selfish and interested in reaching their own optimum slicing policy, and given the RAN infrastructure is shared among MVNOs, slicing policies belonging to the same cluster will necessarily affect each other.

For these reasons, congestion games (CGs) [14]–[16] become the most natural choice to design and analyze efficient and distributed RAN slicing algorithms. Indeed, by using the theory of CGs, we are able to (i) demonstrate the existence and uniqueness of the *Nash equilibrium* (NE) associated to the CG; (ii) prove that its *Price of Anarchy* (PoA) is limited to $\frac{3}{2}$; and (iii) formulate two privacy-preserving distributed algorithms that provably converge to the unique NE.

We now formulate the RAN slicing problem as a congestion game and we define the slicing cost function used by MVNOs to select their optimum policy.

A. Congestion Game Problem Formulation

Let us define and consider the following *weighted* congestion game (CG)

$$\mathcal{G} = \left(\mathcal{M}, (n_m)_{m \in \mathcal{M}}, \mathcal{R}, \mathcal{S}, (c_m)_{m \in \mathcal{M}}\right), \quad (4)$$

where $\mathcal{M}$ is the *player set* (i.e., the set of MVNOs), n_m are the *weights* of the congestion game and represent the expected number of MUs in the cluster for MVNO m, $\mathcal{R}$ is the *resource set*, $\mathcal{S} = \prod_{m \in \mathcal{M}} \mathcal{S}_{m,c}$ is the *strategy space*, and $(c_m)_{m \in \mathcal{M}}$ is the *cost function set*, where c_m will be defined in Section IV-B. A *strategy* for MVNO m consists in the selection of the allocation vector $\boldsymbol{\xi}_m$. Hence, a strategy profile is represented by $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_M)$.

First, note that the set of MVNOs is *finite*, and each MVNO is an atomic decision maker that aims at minimizing the cost associated to its slicing policies. Since the cost experienced by each MVNO depends on the RRHs they select, the congestion game $\mathcal{G}$ has to be weighted and to have *player-specific cost functions*. Furthermore, a slicing policy is derived by allocating different number $\xi_{m,r}$ of MUs to multiple RRHs in the cluster, implying that $\mathcal{G}$ has to be continuous and with *splittable flows* [28]. To summarize, the slicing problem can be modeled as an atomic weighted CG with splittable flows on R parallel links. Figure 2 shows a CG with two MVNOs. In this example,

five RRHs are divided into two distinct RRH classes $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$. Finally, $\xi_{m,r}$ represents the number of MUs allocated to RRHs r by MVNO m.

B. Definition of Slicing Cost Function

RRHs are heterogeneous in terms of type, position and available resources (i.e., spectrum and computational power). Thus, when selecting an RRH r, an MVNO m will incur in a cost that depends on the following two aspects:

1) *RRH Congestion.* RRHs serving an excessive amount of MUs are expected to provide low QoS to their MUs. Instead, RRHs that serve fewer MUs might provide their served MUs with better quality communication. Therefore the RRH congestion should be taken into account. The *congestion level* on RRH r under policy $\boldsymbol{\xi}$ can be measured as follows:

$$\phi_r(\boldsymbol{\xi}) = \frac{1}{N_r} \sum_{m \in \Theta_r(\boldsymbol{\xi})} \xi_{m,r}. \quad (5)$$

where N_r has been defined in Section VIII-A and depends on several RRH-specific parameters such as the resource availability Λ_r at RRH r and the quality of the wireless channel. It is worth noting that the congestion cost not only depends on the allocation policy $\boldsymbol{\xi}_m$ of MVNO m, but it also depends on other MVNOs' allocation policies.

2) *Deployment Cost.* A second cost term is the *deployment cost* which MVNO m pays to the TO to add RRH r to its slice. Such cost depends on the price p_r associated to RRH r. It models the revenue that the TO expects to receive when renting resources to MVNOs.

The above two aspects can be considered by defining the following resource- and MVNO-specific cost function

$$c_{m,r}(\boldsymbol{\xi}) = \phi_r(\boldsymbol{\xi}) + \pi_m^{\mathcal{P}} p_r \quad (6)$$

where the weight $\pi_m^{\mathcal{P}}$ is a resource- and MVNO-specific parameter that weighs the two terms in (6). $\pi_m^{\mathcal{P}}$ can be used to model a wide variety of network scenarios. For example, $\pi_m^{\mathcal{P}} = 0$ can be used to model a MVNO m that cares about congestion only irrespective of the incurred deployment monetary cost.

The total cost incurred by MVNO m under allocation policy $\boldsymbol{\xi}$ can be expressed as

$$c_m(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) = \sum_{r \in \Gamma_m(\boldsymbol{\xi})} \xi_{m,r}\, c_{m,r}(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) \quad (7)$$

$$= c_m^{\mathcal{C}}(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) + c_m^{\mathcal{P}}(\boldsymbol{\xi}_m) \quad (8)$$

where

$$c_m^{\mathcal{C}}(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) = \sum_{r \in \Gamma_m(\boldsymbol{\xi}_m)} \xi_{m,c,r}\, \phi_r(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) \quad (9)$$

$$c_m^{\mathcal{P}}(\boldsymbol{\xi}_m) = \pi_m^{\mathcal{P}} \sum_{r \in \Gamma_m(\boldsymbol{\xi}_m)} \xi_{m,r}\, p_r \quad (10)$$

In the expressions above, (9) measures the overall congestion level experienced by all MUs served by the MVNO m under the slicing policy $\boldsymbol{\xi} = (\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m})$; and (10) represents the total monetary deployment cost of the requested slice.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

D'ORO *et al.*: LOW-COMPLEXITY DISTRIBUTED RADIO ACCESS NETWORK SLICING

5

V. NASH EQUILIBRIUM ANALYSIS

In the previous section, we showed how the slicing problem can be modeled as a congestion game $\mathcal{G}$. In this section, we solve $\mathcal{G}$ by computing the Nash Equilibrium (NE) of the game, defined as follows:

Definition 1: A strategy profile $\boldsymbol{\xi} = (\boldsymbol{\xi}_m)_{m \in \mathcal{M}} \in \mathcal{S}$ is a NE for the congestion game $\mathcal{G}$ if, for all $m \in \mathcal{M}$, $(\boldsymbol{\xi}'_m, \boldsymbol{\xi}_{-m}) \in \mathcal{S}$ it holds that $c_m(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) \leq c_m(\boldsymbol{\xi}'_m, \boldsymbol{\xi}_{-m})$, where $c_m(\cdot)$ is the cost function defined in (7).

Thus, a NE is a strategy profile such that no MVNO can further reduce its own cost function by unilaterally switching to another strategy.

In the following, we will focus our efforts to address the next four questions:

1) Does game $\mathcal{G}$ admit one or more NEs?
2) Is the NE efficient with respect to the optimal solution?
3) Can we design an efficient distributed algorithm that will help the MVNOs reach the NE without interacting with each other?

Question 1) will be answered in Section V-A, while questions 2) and 3) will be answered in Sections V-B and VI, respectively.

A. Existence and Uniqueness of the NE

We first introduce the concept of *exact potential games* in Definition 2, and then we show that game $\mathcal{G}$ is an exact potential game in Proposition 3.

Definition 2: The game $\mathcal{G}$ is an exact potential game if there exists a function $\Phi : \mathcal{S} \to \mathbb{R}$ such that $c_m(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) - c_m(\boldsymbol{\xi}'_m, \boldsymbol{\xi}_{-m}) = \Phi(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) - \Phi(\boldsymbol{\xi}'_m, \boldsymbol{\xi}_{-m})$ for all $m \in \mathcal{M}$, $\boldsymbol{\xi}_m, \boldsymbol{\xi}'_m \in \mathcal{S}_m$, and $\boldsymbol{\xi}_{-m} \in \mathcal{S}_{-m}$, where $\mathcal{S}_{-m} = \prod_{k \in \mathcal{M}, k \neq m} \mathcal{S}_k$.

For each MVNO $m \in \mathcal{M}$ and RRH $r \in \mathcal{R}$, let $\eta_{m,r}$ be defined as follows:

$$\eta_{m,r} = \pi_m^{\mathcal{P}} p_r \qquad (11)$$

Proposition 3 (Potential Function and NE Uniqueness): The congestion game $\mathcal{G}$ is an exact potential game with potential function

$$\Phi(\boldsymbol{\xi}) = \sum_{r \in \mathcal{R}} \sum_{m \in \mathcal{M}} \left(\frac{1}{N_r} \xi_{m,r}^2 + \eta_{m,r} \xi_{m,r} \right.$$
$$\left. + \frac{1}{N_r} \sum_{k < m} \xi_{m,r} \xi_{k,r} \right). \qquad (12)$$

Furthermore, the congestion game $\mathcal{G}$ admits a unique NE.

For a detailed proof see Appendix A. Instead, for a physical interpretation of the potential function (12) see Appendix B.

It is worth mentioning that the potential function in (12) depends on the player and resource-specific parameters $\eta_{m,r}$. Accordingly, it differs from potential functions for atomic congestion games derived in other papers such as [29] that considers resource-specific cost functions only. We would also like to strengthen the fact that having player and resource-specific cost functions not only makes our model general, but it also makes it possible to model all those cases where MVNOs can differentiate among the available RRHs according to the presence of users in a given area, their number and required QoS.

B. The Price of Anarchy (PoA)

In this section, we investigate the efficiency of the unique NE of the CG. Specifically, we will demonstrate that the *Price of Anarchy* (PoA) of the game is a small number. In particular, the PoA is a metric which measures the efficiency of the NE w.r.t. a social optimum solution, where the latter is computed as follows:

$$\boldsymbol{\xi}^{\mathrm{OPT}} = \arg \min_{\boldsymbol{\xi} \in \mathcal{S}} \sum_{m \in \mathcal{M}} c_m(\boldsymbol{\xi}) \qquad (13)$$

In our considered case, let $\boldsymbol{\xi}^{\mathrm{NE}}$ be the unique NE of $\mathcal{G}$, the PoA can be expressed as

$$\mathrm{PoA} = \frac{\sum_{m \in \mathcal{M}} c_m(\boldsymbol{\xi}^{\mathrm{NE}})}{\sum_{m \in \mathcal{M}} c_m(\boldsymbol{\xi}^{\mathrm{OPT}})} = \frac{C^{\mathrm{NE}}}{C^{\mathrm{OPT}}} \qquad (14)$$

Since each MVNO aims at minimizing its cost function, it holds that $\mathrm{PoA} \geq 1$. Specifically, $\mathrm{PoA} = 1$ ensures the global optimality of the NE, instead, $\mathrm{PoA} >> 1$ indicate poor efficiency of the NE. Even though many congestion games have been shown to have unbounded PoA, i.e., $\mathrm{PoA} = +\infty$, in this work we show that the PoA is upper-bounded by $3/2$. Specifically, in Proposition 4 we first derive a general result on the PoA, then in Theorem 5 we derive a closed-form upper-bound which only depends on the number M of MVNOs.

In line with [30], let β be defined as follows

$$\beta = \sup_{r \in \mathcal{R}} \sup_{\boldsymbol{\xi}, \boldsymbol{\nu} \in \mathcal{S}}$$
$$\left\{ \frac{\sum\limits_{m \in \mathcal{M}} \xi_{m,r}(c_{m,r}(\boldsymbol{\xi}) - v_{m,r}(\boldsymbol{\xi})) + \nu_{m,r}(v_{m,r}(\boldsymbol{\xi}) - c_{m,r}(\boldsymbol{\nu}))}{\sum\limits_{m \in \mathcal{M}} \xi_{m,r} c_{m,r}(\boldsymbol{\xi})} \right\} \qquad (15)$$

where $v_{m,r}$ is the first-order partial derivative of $\xi_{m,r} c_{m,r}$ in (6), and is defined as

$$v_{m,r}(\boldsymbol{\xi}) = c_{m,r}(\boldsymbol{\xi}) + \frac{\xi_{m,r}}{N_r}. \qquad (16)$$

Proposition 4: The relationship $C^{\mathrm{NE}} \leq \frac{1}{1-\beta} C^{\mathrm{OPT}}$ holds, i.e., $\mathrm{PoA} \leq \frac{1}{1-\beta}$.
Please refer to Appendix C for the complete proof.

Theorem 5 (The PoA is upper bounded): The PoA is upper bounded by $\mathrm{PoA}(M) = \frac{3M+1}{2M+2}$. That is, the PoA increases with the number M of MVNOs, and is upper-bounded by $\mathrm{PoA}(+\infty) = 3/2$.

For the detailed proof see Appendix D.

It is worth noting that the result in Theorem 5 matches the result in [31], where it has been shown that the PoA for atomic splittable congestion games with resource-specific cost functions is bounded by $\frac{3M+1}{2M+2}$. Our proof of Theorem 5 follows the same logic behind the proof derived by Cominetti *et al.* [31]. However, we would like to mention that the result in [31] applies to the case where cost functions are resource-specific only, i.e., $c_{m,r} = c_r$ for all $m \in \mathcal{M}$. In Theorem 5, instead, we consider cost functions that account for both player and resource-specific terms such as those in (6) and (11). It follows that it is not possible to readily apply [31] to our case, and though we follow the same rational behind the proof of Cominetti et al., our proof actually extends their results to the more general case of player and resource-specific linear cost functions.

VI. Distributed Algorithms for Computing Nash Equilibrium

In this section, we propose two algorithms that provably converge to the unique NE of game $\mathcal{G}$. Specifically, in Section VI-.1 we develop a *best response* approach to develop a mechanism that computes a NE by solving a convex quadratic programming (QP) problem. Section VI-.2 will be devoted to the design of a learning mechanism which provably converges towards a NE and further reduces the complexity of the NE computation.

1) Best Response Approach: We introduce the concept of *best response* (BR) functions and then prove that algorithms based on BR converge towards the unique NE.

The BR is a function that minimizes the cost of each player m given the strategies $\boldsymbol{\xi}_{-m}$ of the other players. Specifically, the BR for player m is defined as

$$\boldsymbol{\xi}_m^{\mathrm{BR}} = \arg \min_{\boldsymbol{\xi}_m \in \mathcal{S}_m} c_m(\boldsymbol{\xi}_m, \boldsymbol{\xi}_{-m}) \qquad (17)$$

An iterative algorithm where each player updates its current strategy according to (17) is called a *best response dynamics* (BRD). The following Proposition 6 holds.

Proposition 6: The sequential BRD converges towards the unique NE of $\mathcal{G}$. Furthermore, the BR of each MVNO $m \in \mathcal{M}$ is unique.

Proof: Since $\mathcal{G}$ admits a potential function, from (12) we have that any BR that minimizes the cost function of a given player also reduces the value of the potential Φ. Therefore, since the potential is bounded and always non-negative, the sequential BRD is an *improvement path* [32] and will surely converge towards the unique NE. Finally, let us note that the cost functions in (7) are strictly convex in each player's strategy $\boldsymbol{\xi}_m$. Thus, the BR of each MVNO is unique. ∎

Note that Proposition 6 does not guarantee that the convergence towards the unique NE is attained in a finite number of iterations. However, potential games with continuous strategy space possess the *approximate finite improvement path* (A-FIP) property, which ensures the convergence to an approximate NE in a finite amount of iterations. Even though we cannot guarantee the convergence in finite time to the NE, in Section VIII we will show that such convergence is attained in a limited number of iterations in many practical scenarios.

Proposition 6 shows that it is possible to converge towards the NE through iterative BRs. The BR in (17) for any MVNO m and any adversarial strategy profile $\boldsymbol{\xi}_{-m}$ is obtained by solving

$$\min_{\boldsymbol{\xi}_m \in \mathcal{S}_m} \quad \mathbf{f}_m \boldsymbol{\xi}_m + \frac{1}{2} \boldsymbol{\xi}_m^T \mathbf{Q} \boldsymbol{\xi}_m$$
$$\text{subject to} \sum_{r \in \mathcal{R}} \xi_{m,r} = n_m, \ \boldsymbol{\xi}_m \geq 0, \qquad (18)$$

where $\mathbf{f}_m = (f_{m,i})_{i \in \mathcal{M}}$ and $\mathbf{Q} = \mathrm{diag}\left(\frac{2}{N_r}\right)_{r \in \mathcal{R}}$ are a M-dimensional row vector and a $M \times M$ diagonal matrix, respectively. with

$$f_{m,r} = \frac{1}{N_r}\left[\phi_r(\boldsymbol{\xi}) - \xi_{m,r}\right] + \eta_{m,r} \qquad (19)$$

and $\phi_r(\boldsymbol{\xi})$ and $\eta_{m,r}$ being defined in (5) and (11), respectively.

An algorithmic implementation of the sequential BRD is described in Algorithm 1.

Algorithm 1 Sequential BRD
___
1: `Input` $\mathcal{R}; \{(x_r, y_r), p_r, \phi_r(\boldsymbol{\xi})\}_{r \in \mathcal{R}};$ `Output` The unique
   NE of $\mathcal{G}$;
2: **while** Convergence is not achieved **do**
3:     **for each** $m \in \mathcal{M}$ **do sequentially**
4:       $\boldsymbol{\xi}_m = (\xi_{m,r})_{r \in \mathcal{R}} \leftarrow$ the unique solution of Problem (18);
5:     **end for**
6: **end while**
___

Problem (18) is a strictly convex quadratic problem with linear constraints. Accordingly, a solution can be computed in polynomial time (typically in $\mathcal{O}(R^3)$, where R is the number of available RRHs). Furthermore, if the congestion level $\phi_r(\boldsymbol{\xi})$ is publicly available, (19) guarantees that Problem (18) can be locally solved by each MVNO without any communication with the other MVNOs. Accordingly, Algorithm 1 computes a NE in a distributed fashion.

2) A Low-Complexity Learning Approach: In the previous section, we have shown that the BR of each MVNO can be computed by solving a QP problem, which is generally solved in polynomial time $\approx \mathcal{O}(R^3)$. However, if the number R of RRHs is large, to compute a solution to Problem (18) would anyway require a non-negligible amount of time. Motivated by this latter discussion, in this section we provide another approach that builds on learning theory to provide a low-complexity approach to compute the unique NE of $\mathcal{G}$.

Let us introduce the following *exponential learning scheme*

$$\begin{cases} z_{m,r}[n+1] = z_{m,r}[n] - \gamma_n v_{m,r}(\boldsymbol{\xi}[n]) \\ \xi_{m,r}[n+1] = n_m \dfrac{e^{z_{m,r}[n+1]}}{\sum_{k \in \mathcal{R}} e^{z_{m,k}[n+1]}} \end{cases} \qquad (20)$$

where γ_n is the *step-size*, $v_{m,r}$ is defined in (16) and n is the iteration indicator. Intuitively, the exponential map in (20) generates values of the allocation variable $\xi_{m,r}$ which always lie on the boundary of the $(R-1)-$simplex $\mathcal{S}_m$. In Proposition 7, we show that (20) converges to the unique NE of $\mathcal{G}$.

Proposition 7: Let $\boldsymbol{\xi}^* \in \mathcal{S}$ be the unique NE of $\mathcal{G}$. If the step-size γ_n are chosen such that $\sum_{n=1}^{+\infty} \gamma_n^2 < \sum_{n=1}^{+\infty} \gamma_n = +\infty$, the exponential learning scheme in (20) always converges towards $\boldsymbol{\xi}^*$ from any interior point $\boldsymbol{\xi}(0) \in \mathcal{S}$. A suitable choice of γ_n is $\gamma_n = \frac{1}{n^\beta}$, with $\beta \in (0.5, 1]$.

Proof: Let us first derive the continuous time version of (20) as follows:

$$\begin{cases} \dot{z}_{m,r} = v_{m,r}(\boldsymbol{\xi}) \\ \xi_{m,r} = n_m \dfrac{e^{z_{m,r}}}{\sum_{k \in \mathcal{R}} e^{z_{m,k}}} \end{cases} \qquad (21)$$

From the strict convexity of c_m, it follows that $v_{m,r}$ is continuously differentiable, and thus Lipschitz-continuous. Since $v_{m,r}$ is always bounded, (21) admits a unique solution for any interior point $\boldsymbol{\xi}(0)$.[2] From (12), we have that the potential function Φ is strictly convex, which implies that Φ is also star-convex with respect to the unique NE $\boldsymbol{\xi}^*$, i.e., $\Phi(\epsilon\boldsymbol{\xi}^* + (1-\epsilon)\boldsymbol{\xi}) \leq \epsilon\Phi(\boldsymbol{\xi}^*) + (1-\epsilon)\Phi(\boldsymbol{\xi})$ for all $\epsilon \in [0,1]$ and $\boldsymbol{\xi} \in \mathcal{S}$. Let $\boldsymbol{\xi}(0)$ be an interior point for $\mathcal{S}$, it is possible to

---

[2]Note that the initialization variable is $\boldsymbol{z}(0) \in \mathbb{R}^{M \times R}$, and any $\boldsymbol{z}(0)$ with finite elements always generates an interior point $\boldsymbol{\xi}(0)$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

D'ORO *et al.*: LOW-COMPLEXITY DISTRIBUTED RADIO ACCESS NETWORK SLICING

7

show [33, Th. 6] that the star-convexity of the potential with respect to the unique NE implies that the KullbackâŁ"Leibler divergence $D_{KL}(\boldsymbol{\xi}^*||\boldsymbol{\xi})$ is a Lyapunov function for (21) and the unique solution of (21) also corresponds to the unique NE $\boldsymbol{\xi}^*$. Now it suffices to apply [33, Th. 4] to show that (20) converges to $\boldsymbol{\xi}^*$ if $\sum_{n=1}^{+\infty} \gamma_n^2 < +\infty$ and $\sum_{n=1}^{+\infty} \gamma_n = +\infty$. Finally, it is straightforward to verify that $\gamma_n = \frac{1}{t^\beta}$ satisfies the above conditions. ∎

The algorithmic implementation of (20) is described in Algorithm 2.

Algorithm 2 Exponential Learning Scheme

1: Input $\mathcal{R}; \{(x_r, y_r), p_r, \phi_r(\boldsymbol{\xi})\}_{r \in \mathcal{R}}$; Output The unique NE of $\mathcal{G}$;
2: Set $(v_{m,r})_{r \in \mathcal{R}, m \in \mathcal{M}} = 0$; $(z_{m,r})_{r \in \mathcal{R}, m \in \mathcal{M}} = 0$;
3: **while** Convergence is not achieved **do**
4:     **for each** $m \in \mathcal{M}$ **do simultaneously**
5:         $\xi_{m,r} \leftarrow n_m \frac{e^{z_{m,r}}}{\sum_{k \in \mathcal{R}} e^{z_{m,k}}}$;
6:         $z_{m,r} \leftarrow z_{m,r} - \gamma_n v_{m,r}(\boldsymbol{\xi})$;
7:     **end for**
8: **end while**

As already discussed at the end of Section VI-.1, Algorithm 2 can be implemented in a distributed fashion by only making the congestion level $\phi_r(\boldsymbol{\xi})$ on each RRH publicly available. Furthermore, in opposition to the BR-based mechanism in Section VI-.1, lines $5, 6$ in Algorithm 2 have complexity $\mathcal{O}(1)$. Since both $\xi_{m,r}$ and $z_{m,r}$ have to be computed for each RRH in $\mathcal{R}$ and MVNO in $\mathcal{M}$, the overall per-iteration computational complexity is $\mathcal{O}(RM)$, i.e., each iteration of Algorithm 2 can be computed in linear time.

In Proposition 7, we have shown that the learning scheme in (20), and thus Algorithm 2, converges to the unique NE of the game if some conditions on the step-size γ_n are satisfied. As an example, a suitable choice of the step-size is $\gamma_n = \frac{1}{n^\beta}$. Though this latter setting guarantees the convergence of the learning scheme, it might generate slow-convergent dynamics. To overcome this issue, it has been shown in [34] that the convergence process can be improved by using fixed values of the step-size. Although the convergence for this latter approach cannot be theoretically proven, using a fixed step-size still allows Algorithm 2 to converge to the NE with high probability. Accordingly, in this paper we will only consider fixed step-size rules, and we refer the interested readers to [34] for a more detailed discussion on the impact of different step-size rules on the convergence of the exponential learning scheme (20).

A. Extensions to Achieve Minimum QoE Levels

Even though the achievable performance of the network certainly depend on the current slicing policy adopted by each MVNO, it is also true that metrics such as throughput and latency are tightly connected to scheduling policies rather than slicing policies. As an example, although a single MVNO receives all of the resources of a given RRH, it does not necessarily imply that MUs experiencing poor channel conditions will be able to enjoy high throughput and low latency.

The derivation of optimal scheduling policies for the MVNOs is out of the scope of this paper, but in this section we derive intuitive and practical approaches that tackles the minimum QoE problem with a different perspective. Specifically, we identify the two following approaches:

Proportional-Time Slicing: this approach uses *time-sharing* to divide a *slicing window* T_r into multiple temporal slots each of duration

$$\tau_{m,r} = \frac{\xi^*_{m,r}}{\sum_{l \in \mathcal{M}} \xi^*_{l,r}} T_r \qquad (22)$$

where $\boldsymbol{\xi}^* = (\xi^*_{m,r})_{m \in \mathcal{M}, \ r \in \mathcal{R}}$ is the NE computed through Algorithms 1 or 2, and $\sum_{m \in \mathcal{M}} \tau_{m,r} = T_r$ for all $r \in \mathcal{R}$. Within each temporal slot $\tau_{m,r}$, all the Λ_r resources are exclusively allocated to m. Since m owns all the available resources in the corresponding temporal slot, it can derive optimal scheduling policies that meet certain performance metrics.

Barrier-Based Slicing: this approach the minimum QoE requirement is translated into a constraint of the RAN slicing problem, and it is then modeled as barrier-function that generates a cost when the constraint is not satisfied. Let $\sigma^{(min)}_{m,r}$ be the minimum amount of resources that MVNO m is willing to obtain on RRH r to guarantee a minimum QoE level, and let $\boldsymbol{\sigma}^{(min)} = (\sigma^{(min)}_{m,r})_{m \in \mathcal{M}, r \in \mathcal{R}}$. Note that not all realizations of $\boldsymbol{\sigma}^{(min)}$ are feasible. Specifically, $\boldsymbol{\sigma}^{(min)}$ is unfeasible if $\exists r \in \mathcal{R}$ such that $\sum_{m \in \mathcal{M}} \sigma^{(min)}_{m,r} > \Lambda_r$. Furthermore, it is straightforward to show that the condition $\sum_{m \in \mathcal{M}} \sigma^{(min)}_{m,r} > \Lambda_r$ is not sufficient to guarantee the existence of an allocation policy $\boldsymbol{\xi} \in \mathcal{S}$ that satisfies all MVNO' request. For this reason, we define the QoE feasibility set as $\mathcal{F} = \{\boldsymbol{\sigma}^{(min)} : \sum_{m \in \mathcal{M}} \sigma^{(min)}_{m,r} \leq \Lambda_r \wedge \exists \ \boldsymbol{\xi} \in \mathcal{S} \text{ with } \sigma_{m,r} \geq \sigma^{(min)}_{m,r}, \forall m \in \mathcal{M}, \ r \in \mathcal{R}\}$. Under the above assumptions, we consider the perturbed cost function

$$\tilde{c}_m(\boldsymbol{\xi}) = c_m(\boldsymbol{\xi}) + \sum_{r \in \mathcal{R}} \lambda_{m,r}(k_{m,r}\phi_r(\boldsymbol{\xi}) - \xi_{m,r}) \qquad (23)$$

where $\lambda_{m,r} > 0$ is the slope of the barrier and $k_{m,r} = \frac{N_r \sigma^{(min)}_{m,r}}{\Lambda_r}$. When $\boldsymbol{\sigma}^{(min)} \in \mathcal{F}$, this formulation introduces separable linear cost functions, and it is possible to use the same steps in Proposition 3 to derive an exact potential function. Similarly, it can be shown that the slicing game with barrier functions admits a unique NE that can be computed through Algorithms 1 and 2.

Remark 1: The proportional-time slicing reserves all the available resources to each MVNO for a limited amount of time, and thus leaves to the MVNOs the computation of scheduling policies that guarantee minimum QoE requirements. The barrier-based approach, instead, first require the execution of a feasibility test of $\boldsymbol{\sigma}^{(min)}$, then needs a proper tuning of $\lambda_{m,r}$. In general, the higher the value of $\lambda_{m,r}$, the higher the cost experienced by MVNOs when the QoE constraint is not met, the higher the probability that solutions satisfying the QoE requirement are generated. However, if not properly tuned, the $\lambda_{m,r}$ parameter might negatively bias the slicing algorithm by pushing MVNOs towards conservative policies that aims at meeting the constraints only rather than minimizing congestion and deployment costs. To understand the effectiveness of this latter approach, we assess its performance in Fig. 14.

VII. PRICING POLICIES

In the previous sections, we have shown that the developed algorithm does not require perfect knowledge with respect

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

8   IEEE/ACM TRANSACTIONS ON NETWORKING

to MVNOs' parameters, and effectively preserves the privacy of all MVNOs. In addition, although the proposed algorithm has limited complexity, we have proven that close-to-optimal solutions can be obtained with no need to run complex and time-consuming centralized approaches at the TO side. Despite the above important properties, another relevant issue that has not yet been considered is related to pricing policies enforced by the TO and their impact on the achievable profit.

It is worth noting that the set of MVNOs, the availability of RRHs and the position and number of MUs in each cluster vary at each slot t. In such a dynamic scenario, it is straightforward to show that static pricing policies could fail to provide high profit to the TO. On the contrary, adaptive pricing policies are more suited to deal with time-varying systems such as the one we are considering in this paper. For this reason, in this section we focus on adaptive pricing policies, and we propose a stochastic-based algorithm that exploits past observations to adapt the pricing policy at future slots.

Let us focus on a given cluster c. Furthermore, let $\boldsymbol{\xi}^*(t) = (\xi^*_{m,r}(t))_{m \in \mathcal{M}, r \in \mathcal{R}}$ be the unique NE of $\mathcal{G}$ at slot t, and let $n_r(t)$ be the number of MUs served through RRH r at the NE. We have that

$$n_r(t) = \sum_{m \in \mathcal{M}} \xi^*_{m,r}(t). \tag{24}$$

The profit of the TO is defined as

$$\Pi(t) = \sum_{r \in \mathcal{R}} p_r(t) n_r(t) - C(n_r(t)), \tag{25}$$

where $C(n_r(t))$ is the cost experienced by the TO to manage $n_r(t)$ MUs at RRH r.

The allocation variables $\xi^*_{m,r}$ are obtained by iteratively executing either Algorithm 1 or Algorithm 2. Unfortunately, there is no closed-form for those variables, which makes it hard to predict the behavior of the MVNOs, and the actual allocation of the MUs. Also, $\xi^*_{m,r}$ depends on the weighting parameters $\pi^{\mathcal{P}}_m$. Those values are not known by the TO and might vary at each slot t. Therefore, an adaptive pricing policy should be considered such that the price p_r can be updated at each slot t.

In this paper, we propose the following stochastic-based approximation pricing mechanism

$$p_r(t+1) = p_r(t) + \sigma \left[n_r(t) - n_r(t-1) \right], \tag{26}$$

where $\sigma > 0$ is a fixed step-size used to weigh the two terms in the stochastic procedure. The proposed pricing scheme in (26) works as follows. If $n_r(t) > n_r(t-1)$, i.e., the number n_r of MUs served through RRH r has been increased in the last slot, then the price at the next slot is increased as well, i.e., $p_r(t+1) > p_r(t)$. Otherwise, if $n_r(t) < n_r(t-1)$, the price is decreased, and $p_r(t+1) < p_r(t)$.

Furthermore, since the achieved profit must always be non-negative, the condition $p_r(t+1) \geq C(n_r(t))$ has to be satisfied for each RRH r. Since the stochastic procedure in (26) might generate a price $p_r(t+1)$ that violates the above constraint, a minimum value $p_r(t+1) = C(n_r(t))$ is considered at each iteration of (26).

VIII. Performance Evaluation

In this section, we investigate on the performance of the proposed congestion game-based slicing mechanism through numerical simulation and experimental results. To emulate a
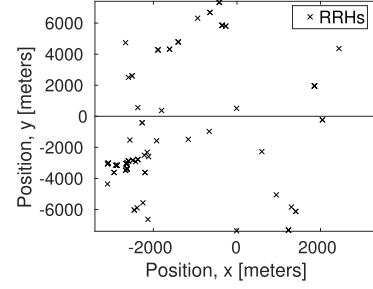

Fig. 3. The considered cluster with the position of 100 RRHs taken from the OpenCellID database.

realistic network setup, in our simulations we have extracted a cluster $\mathcal{R}^{(\text{FULL})}$ of multiple RRHs deployed in Boston, MA (USA) from the OpenCellID database [35]. A cluster consisting of 100 RRHs, and the position of each RRH is illustrated in Fig. 3. We assume that the density of wireless devices in the considered cluster is set to 5000 devices/km^2. Furthermore, we consider the case where a set $\mathcal{M}$ of $M = 20$ MVNOs are willing to competitively deploy RAN slices to provide MUs with wireless access to the internet. We assume that each MVNO m serves the same number n_m of devices. Hence, the density of wireless devices served by each MVNO is $5000/M$ devices/km^2.

We assume that all RRHs are LTE base stations operating at 2.4GHz. The path-loss exponent is set to $\alpha = 3$, e.g., urban scenario, and the channel noise power is $N = -174$dBm/Hz [34]. We consider isotropic antennas with antenna gain equal to 3dBi, reference distance $d_0 = 1$m [36] and $k = 9.89 \cdot 10^{-5}$ [37].

We assume that the amount of available resources at each RRH r, i.e., Λ_r, is represented by the number of LTE Resource Elements (RE) within an LTE Resource Block (RB). Specifically, we assume that each RB consists of $s_r = 7$ OFDM symbols transmitted over $c_r = 12$ subcarriers for all $r \in \mathcal{R}$. Note that the number of available RBs, and thus Λ_r depends on the LTE channel bandwidth. Specifically, we have that $\Lambda_r = N_{\text{RB}} \cdot s_r \cdot c_r$, where $N_{\text{RB}} \in \{25, 50, 100\}$ represents the number of available RBs in the LTE system when the bandwidth is set to 5, 10, and 20 MHz, respectively. Unless otherwise stated, we assume $b = 20$MHz. That is, $\Lambda_r = 100$ REs are available for MU transmission at each RRH. Furthermore, to investigate the impact of different minimum Quality of Experience (QoE) levels on the performance of the slicing mechanism, in our simulations we consider three different minimum SINR requirements such that SINR$_r \in \{0, 10, 20\}$ dB.

The price p_r associated to each RRH $r \in \mathcal{R}^{(\text{FULL})}$ is assumed to be generated according to a normal distribution with mean value $\mu_p = 10$ Price Units (PU) and standard deviation $\sigma_p = 4$ PU. Instead, the weight $\pi^{\mathcal{P}}_m$ in (6) and the access rate μ of MUs in (27) are assumed to be uniformly distributed and are randomly generated at each simulation run. More in detail, $\pi^{\mathcal{P}}_m$ takes value in $[0, \pi^{\mathcal{P}}]$, with $\pi^{\mathcal{P}} = 5 \cdot 10^{-4}$, while $\mu \in [0, 1]$.

To investigate the impact of the number R of RRHs deployed in the network on the achievable performance of the network, for any given $R \in \mathbb{Z}$, in our simulations we generate a subset $\mathcal{R} \subseteq \mathcal{R}^{(\text{FULL})}$ of R RRHs which are randomly picked from $\mathcal{R}^{(\text{FULL})}$. The results presented in the following are averaged over 2000 independent simulation runs. The 95%

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

D'ORO *et al.*: LOW-COMPLEXITY DISTRIBUTED RADIO ACCESS NETWORK SLICING 9

confidence intervals are not shown when below 2% of the average.

Before going into the details of our performance evaluation results, in Section VIII-A we consider a case-study that will be useful to our investigation. Specifically, we derive a worst-case estimation of the parameter N_r that ties together minimum SINR requirements, number of resources, user' distribution and number.

A. Evaluating N_r : A Case-Study

Let us consider an MU communicating with RRH r. The received power at the MU's side can be denoted as $p_r = k \left(\frac{d_0}{d_r} \right)^\alpha$, where k is a constant that accounts for transceiver design parameters such as antenna gains, channel characteristics, transmission power of the RRH and operating frequency; d_0 is a reference distance for far-field communication; d_r is the distance between the MU and r; and α is the path loss coefficient [37]. Specifically, $\alpha = 1$ corresponds to the linear path loss model, $\alpha \leq 2$ models the free space or indoor path loss, and $\alpha > 3$ models urban and sub-urban scenarios.

Let us assume that the RRH r serves all MUs in its range with the same transmission power level p_r, and let $\tilde{N}_r$ be the number of MUs in the coverage radius of r. Let us consider a single resource (e.g., one sub-carrier or a RB), the Signal-to-Interference-plus-Noise Ratio (SINR) measured by each MU at distance d_r from r can be written as $\text{SINR}_r = \frac{p_r}{N + \tilde{N}_r p_r \mu_r}$, where N is the channel noise power, and $\mu_r \in [0, 1]$ represents the percentage of MUs currently served by the RRH when the number of MUs within the RRH coverage range is maximum and equal to $\tilde{N}_r - 1$, i.e., the worst-case scenario.

Let $\text{SINR}_r^{(min)}$ be the minimum QoS level that each RRH wants to provide to each MUs. To meet the target SINR level, some straightforward calculations show that the following relationship must hold

$$\tilde{N}_r \leq 1 + \frac{1}{\mu_r} \left(\frac{1}{\text{SINR}_r^{(min)}} - \frac{N}{p_r} \right) = N_r^{(MAX)} \qquad (27)$$

Since the number of resources at RRH r is Λ_r, we have that a worst-case estimation of N_r gives $N_r = N_r^{(MAX)} \Lambda_r$.

Remark 2: We notice that (i) the maximum number of users that can be served by RRH r is proportional to the available network resources Λ_r and decreases as μ_r increases; and that (ii) N_r decreases as the minimum SINR level and the distance d_r increase.

B. PoA and Convergence Speed Analysis

We first investigate the performance of the proposed RAN slicing algorithms as compared to an algorithm computing the social optimum as discussed in Section V-B. In Fig. 4, we show the PoA of the proposed slicing solution as a function of the ratio R/M and for different values of N_{RB} when $\mu = 0.8$. The figure concludes that the Price of Anarchy (PoA) is small and never exceeds 1.08 in all the considered cases, i.e., the proposed solution provides near-optimal slicing of the network resources even though the cluster consists of approximately 100 RRHs. Furthermore, Fig. 4 shows that the proposed solution achieves better performance in terms of optimality gap when a small number of RBs is considered.

Fig. 5 shows the impact of the weighing parameter $\pi_m^{\mathcal{P}}$ on the cost function (7) of the PoA. The weight $\pi_m^{\mathcal{P}}$ is uniformly
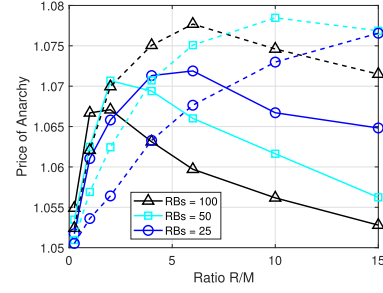

Fig. 4. The PoA of the proposed slicing solution as a function of the ratio R/M for different number N_{RB} of RBs and minimum SINR requirement (Solid lines: SINR = 0dB; Dashed lines: SINR = 20dB).
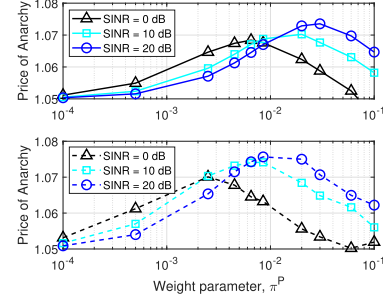

Fig. 5. The PoA of the proposed slicing solution as a function of $\pi_m^{\mathcal{P}}$ for different values of the minimum SINR requirement and number of RRHs (Solid lines: $R = 20$; Dashed lines: $R = 50$).
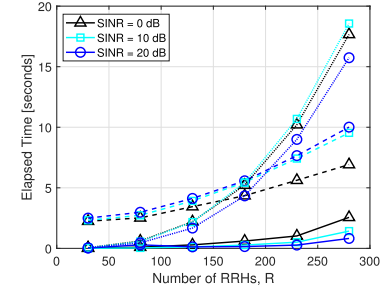

Fig. 6. Execution time of several solutions for different minimum SINR requirements (Solid lines: Algorithm 2; Dashed lines: Algorithm 1; Dotted lines: Centralized optimal solution).

distributed in $[0, \pi^{\mathcal{P}}]$. Accordingly, in Fig. 5 we let the upper-bound $\pi^{\mathcal{P}}$ vary for different values of the minimum SINR requirement. Fig. 5 clearly shows the existence of two distinct regions; the first region is associated to small values of $\pi^{\mathcal{P}}$, i.e., those network scenarios where MVNOs primarily aim at minimizing the congestion cost in (7). In this case, the PoA increases until a maximum value is attained. Instead, when higher values of $\pi^{\mathcal{P}}$ are considered, i.e., the deployment cost associated to each slice is no more negligible, the PoA asymptotically decreases to one until near-optimality is achieved.

Fig. 6 compares the execution time of the solution proposed in this paper. Specifically, we compare the time needed by a centralized algorithm to compute an optimal solution, and that needed by Algorithms 1 and 2 to compute a NE. The results conclude that the centralized algorithm (dotted lines) suffers from severe complexity issues when dense networks are considered. On the contrary, both the BRD (dashed lines) and learning (solid) algorithms show faster convergence rate. Furthermore, Algorithm 1 still requires to solve the QP problem (18), which usually has polynomial computational complexity. Instead, Algorithm 2 has linear per-iteration
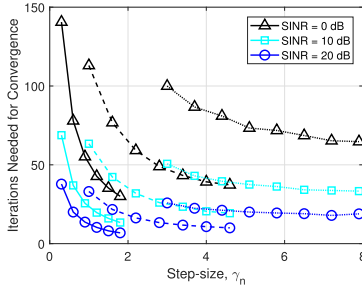
Fig. 7. Number of iterations needed by Algorithm 2 to reach the NE as a function of γ_n for different minimum SINR requirements and number of RRHs (Solid lines: $R = 20$; Dashed lines: $R = 50$; Dotted lines: $R = 100$).
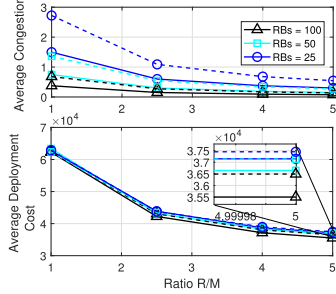

Fig. 8. Average congestion level on each RRH and deployment cost of a slice as a function of the ratio R/M for different number of RBs and minimum SINR requirements (Solid lines: $\mathrm{SINR} = 0\mathrm{dB}$; Dashed lines: $\mathrm{SINR} = 20\mathrm{dB}$).

and per-user complexity, which results in faster convergence towards the NE.

To investigate the convergence speed of Algorithm 2 under the fixed step-size rule, in Fig. 7 is shown the number of iterations needed by the learning algorithm to converge to the NE as a function of the value of the step-size parameter for different minimum SINR requirements and values of the number R of RRHs. It is worth noting that the higher the value of the step-size, the higher the convergence speed of the algorithm. Furthermore, the convergence of Algorithm 2 is slowed down when the number R of RRHs is large. However, by increasing the value of the fixed step-size, the convergence speed is reduced even in those scenarios where the number of RRHs and MVNOs is high.

C. Congestion and Deployment Costs Analysis

Fig. 8 shows the impact of the ratio R/M on both the average congestion and the deployment cost of the network for different values of the number of RBs and the minimum SINR requirement when $\mu = 0.8$ and $\pi^{\mathcal{P}} = 5 \cdot 10^{-2}$. Specifically, the congestion at each RRH is evaluated as in (5). Instead, the deployment cost of each MVNO $m \in \mathcal{M}$ is computed according to (10) and then normalized by the corresponding weight $\pi^{\mathcal{P}}_m$ to obtain the actual monetary cost. The figure concludes that the congestion is decreasing as a function of R/M. Intuitively, this is because the deployment of additional RRHs on the network allows the MVNOs to generate RAN slices which contain different RRHs, thus inevitably reducing the congestion level on those RRHs. Furthermore, it is shown that the deployment cost decreases as well when a larger number of RRHs are deployed. In this case, MVNOs can select those RRHs with a low price p_r, thus reducing the overall monetary expenditure. Fig. 8 also shows that when $\mathrm{SINR} = 20\mathrm{dB}$ (dashed lines), both the congestion level and the deployment cost of the network are higher as compared to the case where
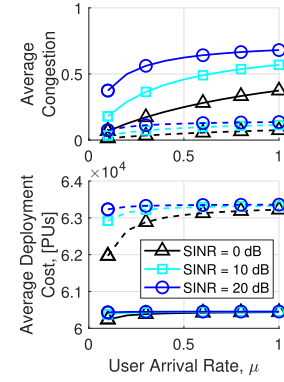

Fig. 9. Average congestion level on each RRH and deployment cost as a function of μ for different values of the minimum SINR requirement and number of RRHs (Solid lines: $R = 20$; Dashed lines: $R = 100$).

$\mathrm{SINR} = 0\mathrm{dB}$ (solid lines). In other words, higher values of the minimum SINR requirements reduce the maximum number of MUs that can be served by each RRH, which, on average, leads to higher values of congestion and deployment cost.

From (27), the number N_r of MUs that can be served with high QoE by each RRH decreases as the access rate μ of MUs increases. As a consequence, the congestion on each RRH increases due to the high traffic demand generated by MUs. This phenomenon is depicted in Fig. 9 where we show the average congestion level on each RRH and deployment cost of a slice as a function of the access rate μ for different values of the minimum SINR requirement and R. Intuitively, the congestion increases when both μ and the minimum SINR level are high, and a small number of RRHs is deployed. Accordingly, to support the high traffic demand, MVNOs add more RRHs to their slice, which increases the deployment cost of each slice.

The impact of $\pi^{\mathcal{P}}_m$ on the congestion level of the sliced network is investigated in Fig. 10, where we show the average congestion level and deployment cost as a function of $\pi^{\mathcal{P}}$ for different values of the minimum SINR requirement and number of RRHs when $\mu = 0.8$. It is worth noting that $\pi^{\mathcal{P}} = 0$ implies that the cost function in (7) does not account for the monetary price of slices anymore. From a purely practical point of view, this scenario relates to the case where MVNOs have a very large (and possibly infinite) monetary budget. Therefore, each MVNO seeks to individually minimize the congestion level of its corresponding slice while disregarding any cost-efficient spending policy. For this reason, the lowest values of congestion are achieved when $\pi^{\mathcal{P}} = 0$. On the contrary, the average congestion increases when $\pi^{\mathcal{P}}$ increase as well. That is, when MVNOs are more cautious about their monetary expenditure, they are prone to select cheap RRHs, even though these RRHs have already been selected by other MVNOs in the network. This is clearly visible in Fig. 10, which shows that the deployment cost of the slices decreases as $\pi^{\mathcal{P}}$ increases. In fact, by increasing the weight parameter, MVNOs are more affected by deployment costs and converge towards more conservative slicing strategies to avoid high monetary expenses. Finally, Fig. 10 shows that the average congestion when only few RRHs are available (i.e., dotted lines) is higher than the case where a high number R of RRHs is deployed on the network (i.e., dashed lines). Intuitively, when more RRHs are available, it is easier for MVNOs to select those RRHs that experience very low-congestion or have not yet been selected by other MVNOs.
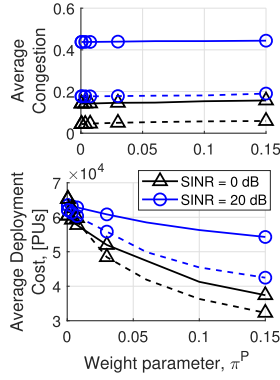
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

D'ORO *et al.*: LOW-COMPLEXITY DISTRIBUTED RADIO ACCESS NETWORK SLICING 11

Fig. 10. Average congestion level on each RRH and deployment cost as a function of $\pi_m^{\mathcal{P}}$ for different values of the minimum SINR requirement and number of RRHs (Solid lines: $R = 20$; Dashed lines: $R = 50$).

D. Profit Analysis

In this section, we investigate the profit achieved by the TO and we compare different pricing policies. Specifically, we consider three different pricing schemes as follows:

- *Uniform Pricing:* all the RRHs in $\mathcal{R}$ are equally priced, i.e., a fixed price equal to $p_r = \mu_p$ is enforced on each RRH $r \in \mathcal{R}$;
- *Weighted Pricing:* the price enforced on each RRH is proportional to N_r. Specifically, the price p_r for the RRH $r \in \mathcal{R}$ is set to $p_r = \mu_p \frac{N_r}{\max\limits_{e \in \mathcal{R}} N_e}$. Under this pricing policy, the TO fixes higher prices to those RRHs which can serve a higher number N_r of devices simultaneously;
- *Adaptive Pricing:* this policy has been introduced and discussed in Section VII. Intuitively, at each slot t, the price of RRH r is updated according to (26) by considering the congestion level on r at slot $t - 1$. From (26), it can be shown that the adaptive scheme will eventually converge towards a stable pricing policy when $t \to \infty$.

The profit achieved by the TO is shown in Fig. 11 as a function of the mean value μ_p for different pricing schemes. Since the proposed adaptive pricing model updates the price of the RRHs at each slicing slot t, in Fig. 11 we show the profit achieved by the TO when $t \to \infty$. As expected, the profit achieved by the TO increases when the value of μ_p increases as well, and it is shown that the proposed pricing model outperforms the other mechanisms.

E. Experimental Results

To understand the impact of our system in real-world deployments and compare the convergence speed of the proposed solution with that achieved by an optimal centralized solution, Fig. 12 shows the experimental results obtained by executing both the learning and the centralized algorithms on a testbed deployed on the Amazon Elastic Computing (EC2) cloud service. Specifically, we have deployed the TO's code on a *t2.micro* instance, having 1 virtual CPU, one 1 GByte of RAM with 3.3GHz clock speed. We have deployed the code implementing the MVNO's slicing policy learning algorithm on an iMac desktop computer located on Northeastern University campus. To provide a fair comparison with a centralized optimal solution As we can see from Fig. 12, the learning algorithm reaches a NE in a limited amount of time (less that 13 seconds for 50 RRHs and 40 MVNOs), despite the overhead due to communication between the TO and the MVNOs.
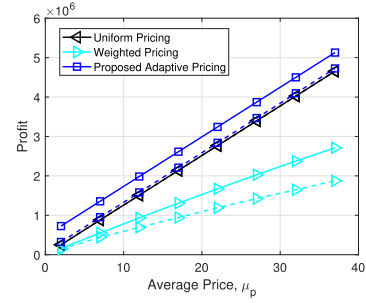

Fig. 11. Profit of the TO as a function of the mean value μ_p for different pricing schemes and values of $\pi_m^{\mathcal{P}}$ (Solid lines: $\pi_m^{\mathcal{P}} = 5 \cdot 10^{-5}$; Dashed lines: $\pi_m^{\mathcal{P}} = 10^{-2}$).
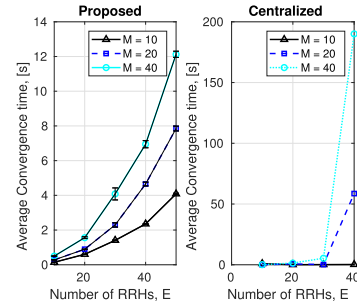

Fig. 12. Experimental results of convergence time as a function of the number R or RRHs for different values of the number M of MVNOs.
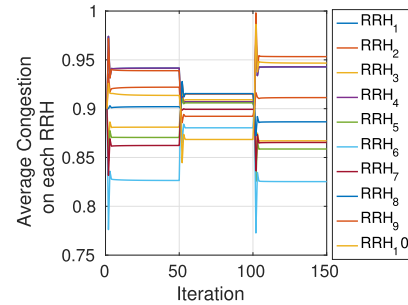

Fig. 13. Dynamic behavior of the proposed slicing algorithm.

It is worth noting that the centralized solution outperforms our solution only for very small network instances, i.e., $M = 10$. In this case, in fact, communication delays play a relevant role in the convergence speed of our solution. On the contrary, our solution is 10 times faster than the centralized one when large network deployments are considered, thus showing the effectiveness of our framework.

In Fig. 13, we show how the proposed RAN slicing algorithm adapts to network changes over time when $M = 3$ and $R = 10$. We assume that the number $n_m(t)$ of MUs served by MVNO m and their position changes over time. Specifically, every 50 iterations of Algorithm 2, we randomly generate a new network configuration by updating both the number of MUs and their expected position in the cluster. Accordingly, Fig. 13 shows the average congestion level on each RRH in the network at each iteration of Algorithm 2. It is worth noting that Algorithm 2 quickly adapts to network changes and rapidly converges towards a NE in few iterations even though the network configuration changes over time.
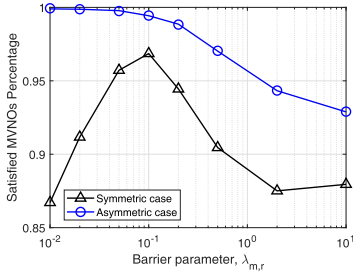
Fig. 14. Percentage of satisfied MVNOs as a function of the barrier parameter $\lambda_{m,r}$.

F. Minimum Requirements Analysis

To assess the performance of the barrier-based slicing approach in Section VI-A, we consider two different cases as follows. A *symmetric case* where all MVNOs request the same amount of resources for each RRH, i.e., $\sigma_{m,r}^{(min)} = \sigma_{n,r}^{(min)} = \Lambda_r/M$ for all $m, n \in \mathcal{M}$ and $r \in \mathcal{R}$. And an *asymmetric case* where we divide the MVNO set into two partitions $\mathcal{M}^I$ and $\mathcal{M}^{II}$ of equal cardinality, and for any $r \in \mathcal{R}$ we set $\sigma_{m,r}^{(min)} = 0.75\Lambda_r/M$ for all $m \in \mathcal{M}^I$, and $\sigma_{m,r}^{(min)} = 0.25\Lambda_r/M$ for all $m \in \mathcal{M}^{II}$. As already discussed in Section VI-A, not all request profiles $\boldsymbol{\sigma}^{(min)}$ admit a feasible solution. Accordingly, we assume that each MVNO activates its minimum resource requests on a subset of cardinality $R/2$ only. In our simulations, the subset of selected RRHs is randomly generated at each simulation run. In Fig. 14, we show the percentage of satisfied MVNOs as a function of the barrier parameter $\lambda_{m,r}$. It is shown that the barrier-based slicing approach is well-suited to provide MVNOs with the required amount of resources. However, while very small values of $\lambda_{m,r}$ should be considered for the asymmetric case, larger values are required for the uniform case.

IX. RELATED WORK

Due to its capability to enable agile and efficient sharing of wireless network resources in 5G systems, the network slicing problem has attracted increasing attention in the literature [10], [13], [38]–[41]. Network slicing is not trivial, and it can be shown that in general it is NP-hard [10]. Centralized solutions have been proposed for both the slicing of both the backbone and RAN portions of the network [10], [13], [42]–[46], but they generally require perfect knowledge of the network configuration and suffers from high complexity issues. As a consequence, a variety of low-complexity heuristic and distributed approaches have been proposed [47]–[51]. However, the above solutions do not consider the adversarial behavior of virtual operators, which makes it hard to implement them in competitive scenarios where privacy must be preserved among multiple entities.

It is worth noting that many instances of the network slicing problem can actually be reduced to the so-called Virtual Network Embedding (VNE) problem [52], which consists in the problem of generating a set of virtual networks upon a shared physical infrastructure while satisfying a given set of constraints. Unfortunately, the VNE problem is NP-hard. Thus, to design effective algorithms with low computational complexity, a variety of centralized and distributed heuristic approaches have been proposed in the literature [52]. However,

all of those solutions are not well-suited to deal with the competitive slicing of the RAN, where information on the expected number of users and their distribution in the network is available. Thus, they cannot be straightforwardly applied to the RAN slicing problem we are tackling in this paper.

Competition among virtual operators has been recently considered in [27], where game-theoretical tools have been effectively exploited to address the RAN slicing problem. In [27], a RAN slicing game is developed, where the available RAN resources are allocated to the MUs through an auction-based mechanism which competitively maximizes the data rate of each slice while guaranteeing a certain level of fairness.

Though the problem tackled in [27] is similar to that addressed in this paper, there are several fundamental differences. First, in [27] a fixed share of RAN resources is assigned to each MVNO a-priori, and the auction game is then played by the MVNOs to decide on how to divide those resources to the mobile users. In our paper, conversely, the share of RAN resources allocated to each MVNO has been obtained as a consequence of a non-cooperative game played among the competing MVNOs. Another important difference is that perfect knowledge about the presence of users is assumed in [27]. In our paper, instead, we have relaxed this assumption by assuming that only statistical information on the presence and position of the users is available.

In this paper, we have leveraged on a particular class of congestion games, i.e., the atomic splittable congestion games. Although the uniqueness of the NE for this class of games is a well-known result in the literature, how to efficiently compute a NE is still a challenging issue. Closed-form solutions have been derived in [53], which however only hold for resource-specific cost functions. When closed-form solutions cannot be derived, then traditional sequential BRD have been considered [54]. However, those approaches generally require to solve QP problems, and are guaranteed to converge to the unique NE only after an infinite number of iterations. Recently, a quantization approach which can compute the NE in a finite amount of time has been proposed in [28]. However, it suffers from highly polynomial complexity, which makes it hard to implement it in dynamic scenarios. Upper-bounds on the PoA of atomic congestion games have been investigated in [31] and [55], but they only apply to the case of resource-specific cost functions. Similarly, upper-bounds on the PoA for a particular class of congestion games have been derived in [56]. Specifically, authors have shown that it is possible to derive an even tighter bound on the PoA of $4/3$. Although such result is of extreme importance for the game-theory community, [56] does not consider the challenging case where cost functions are both player and resource-specific. Also, the model considered in [56] does not require solutions to lie on the simplex, thus preventing the use of such a result for the problem we considered in this paper.

When compared to the above literature on atomic congestion games, in this work we have proposed a simple low-complexity learning algorithm where each MVNO iteratively updates its strategy to provably converge to the unique NE. We have further derived an upper-bound on the PoA of the class of atomic splittable congestion games where resource and player-specific cost functions are considered, thus extending the results in [31] and [55]. Specifically, we have proved that the PoA for this particular class of cost functions matches the PoA of atomic splittable congestion games where only resource-specific cost functions are considered [31].

X. Conclusions

In this paper, we have leveraged congestion games and learning mechanisms to design a distributed solution to the wireless network slicing problem. The proposed solution accounts for the limited availability of wireless resources and considers several aspects such as congestion, deployment costs and distance among the RRHs and MUs. We have shown that the proposed solution achieves near-optimal global performance with limited computational complexity while preserving the privacy of MVNOs. Numerical and experimental results have shown the effectiveness of the proposed approach in terms of cost reduction, scalability and convergence time.
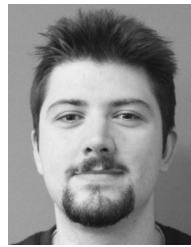
References

[1] Ericsson Mobility Report. (2017). *Mobile Subscriptions Outlook.* [Online]. Available: https://www.ericsson.com/en/mobility-report/mobile-subscriptions-worldwide-outlook

[2] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5G: The next generation of mobile communication," *Phys. Commun.*, vol. 18, pp. 64–84, Mar. 2016.

[3] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 106–112, Apr. 2014.

[4] M. R. Palattella *et al.*, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[5] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[6] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[7] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[8] T. Taleb, B. Mada, M. Corici, A. Nakao, and H. Flinck, "PERMIT: Network slicing for personalized 5G mobile telecommunications," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 88–93, May 2017.

[9] N. Nikaein *et al.*, "Network store: Exploring slicing in future 5G networks," in *Proc. 10th Int. Workshop Mobility Evolving Internet Archit.*, 2015, pp. 8–13.

[10] S. Vassilaras *et al.*, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, Aug. 2017.

[11] K. Samdanis, X. C. Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[12] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: Enabling enterprises' own software-defined cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 146–153, Jul. 2016.

[13] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr., Electron. Syst. Eng. (ICAEES)*, Nov. 2016, pp. 414–420.

[14] C. Tekin, M. Liu, R. Southwell, J. Huang, and S. H. A. Ahmad, "Atomic congestion games on graphs and their applications in networking," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1541–1552, Oct. 2012.

[15] S. D'Oro, L. Galluccio, S. Palazzo, and G. Schembra, "Exploiting congestion games to achieve distributed service chaining in NFV networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 407–420, Feb. 2017.

[16] Z. Han, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[17] Amazon Inc. (2017). *Amazon Elastic Compute Cloud (Amazon EC2).* [Online]. Available: https://aws.amazon.com/ec2/

[18] Unwired Labs. (2017). *Opencellid: The World's Largest Open Database of Cell Towers*. [Online]. Available: https://opencellid.org/#zoom=16&lat=37.77889&lon=-122.41942

[19] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

[20] J. Zyren and W. McCoy, "Overview of the 3GPP long term evolution physical layer," Freescale Semicond., Inc., Austin, TX, USA, White Paper 3GPPEVOLUTIONWP, 2007.

[21] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.

[22] *ELTE2.2 DBS3900 LTE FDD Configuration Principles*, Huawei, Shenzhen, China, 2014.

[23] H. D. Trinh, N. Bui, J. Widmer, L. Giupponi, and P. Dini, "Analysis and modeling of mobile traffic using real traces," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.

[24] S. Wang, W. Guo, and M. D. McDonnell, "Distance distributions for real cellular networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr./May 2014, pp. 181–182.

[25] I. F. Akyildiz and W. Wang, "The predictive user mobility profile framework for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 6, pp. 1021–1035, Dec. 2004.

[26] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1790–1821, 3rd Quart., 2017.

[27] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant networks," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[28] T. Harks and V. Timmermans, "Equilibrium computation in atomic split-table singleton congestion games," in *Proc. Int. Conf. Integer Program. Combinat. Optim.* Cham, Switzerland: Springer, 2017, pp. 442–454.

[29] S. Sorin and C. Wan, "Finite composite games: Equilibria and dynamics," *J. Dyn. Games*, vol. 3, no. 1, pp. 101–120, 2016.

[30] T. Roughgarden, "Selfish routing with atomic players," in *Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2005, pp. 1184–1185.

[31] R. Cominetti, J. R. Correa, and N. E. Stier-Moses, "The impact of oligopolistic competition in networks," *Oper. Res.*, vol. 57, no. 6, pp. 1421–1437, 2009.

[32] D. Monderer and L. S. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, no. 1, pp. 124–143, 1996.

[33] P. Mertikopoulos, E. V. Belmega, A. L. Moustakas, and S. Lasaulce, "Distributed learning policies for power allocation in multiple access channels," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 96–106, Jan. 2012.

[34] S. D'Oro, P. Mertikopoulos, A. L. Moustakas, and S. Palazzo, "Interference-based pricing for opportunistic multicarrier cognitive radio systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6536–6549, Dec. 2015.

[35] *OpenCellID*. Accessed: Jan. 2018. [Online]. Available: https://opencellid.org/

[36] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*. London, U.K.: Pearson, 2010.

[37] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[38] A. Gudipati, L. E. Li, and S. Katti, "Radiovisor: A slicing plane for radio access networks," in *Proc. 3rd Workshop Hot Topics Softw. Defined Netw.*, 2014, pp. 237–238.

[39] I. da Silva *et al.*, "Impact of network slicing on 5G radio access networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2016, pp. 153–157.

[40] L. Cano, A. Capone, G. Carello, and M. Cesana, "Evaluating the performance of infrastructure sharing in mobile radio networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 3222–3227.

[41] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Serv. Manage.*, vol. 13, no. 3, pp. 462–476, Sep. 2016.

[42] C. Gao, C. Ozcan, J. Tang, M. C. Gursoy, and W. Zhang, "R-cloud: A cloud framework for enabling radio-as-a-service over a wireless substrate," in *Proc. 24th Int. Conf. Netw. Protocols (ICNP)*, Nov. 2016, pp. 1–10.

[43] C. Liang and F. R. Yu, "Mobile virtual network admission control and resource allocation for wireless network virtualization: A robust optimization approach," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[44] I. Malanchini, S. Valentin, and O. Aydin, "Generalized resource sharing for multiple operators in cellular wireless networks," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Aug. 2014, pp. 803–808.

[45] R. Mahindra, M. A. Khojastepour, H. Zhang, and S. Rangarajan, "Radio access network sharing in cellular networks," in *Proc. 21st IEEE Int. Conf. Netw. Protocols (ICNP)*, Oct. 2013, pp. 1–10.

[46] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[47] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "Cellslice: Cellular wireless resource slicing for active ran sharing," in *Proc. 15th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2013, pp. 1–10.

[48] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *Proc. 80th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.

[49] A. Nakao *et al.*, "End-to-End network slicing for 5G mobile networks," *J. Inf. Process.*, vol. 25, pp. 153–163, Feb. 2017.

[50] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems," in *Proc. 22nd Eur. Wireless Conf. (VDE)*, May 2016, pp. 1–6.

[51] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Multitenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[52] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1888–1906, 4th Quart., 2013.

[53] E. Altman, T. Basar, T. Jiménez, and N. Shimkin, "Competitive routing in networks with polynomial costs," *IEEE Trans. Autom. Control*, vol. 47, no. 1, pp. 92–96, Jan. 2002.

[54] M. Hoefer and A. Skopalik, "Altruism in atomic congestion games," *ACM Trans. Econ. Comput.*, vol. 1, no. 4, 2013, Art. no. 21.

[55] T. Roughgarden and F. Schoppmann, "Local smoothness and the price of anarchy in splittable congestion games," *J. Econ. Theory*, vol. 156, pp. 317–342, Mar. 2015.

[56] R. Johari and J. N. Tsitsiklis, "Efficiency loss in a network resource allocation game," *Math. Oper. Res.*, vol. 29, no. 3, pp. 407–435, 2004.

[57] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica, J. Econ. Soc.*, vol. 33, no. 3, pp. 520–534, 1965.

[58] A. Neyman, "Correlated equilibrium and potential games," *Int. J. Game Theory*, vol. 26, no. 2, pp. 223–227, 1997.

Francesco Restuccia (M'16) received the Ph.D. degree in computer science from the Missouri S&T in 2016, under the supervision of S. K. Das. He has held research positions at The University of Texas at Arlington and the National Research Council of Italy. He is currently an Associate Research Scientist with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. His research interests lie in the theoretical and practical aspects of wireless networked systems, with applications to mobile computing, software-defined radios, and the Internet of Things. He is a member of the IEEE and ACM. He has served on the Technical Program Committee of IEEE LCN (2016–2018), IEEE WoWMoM (2017–2019), IEEE SMARTCOMP (2017), and IEEE INFOCOM (2018). He is a reviewer for several top-level IEEE and ACM conferences and journals.

Tommaso Melodia (M'07–SM'16–F'18) received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2007. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Northeastern University. He is also the Director of Research for the PAWR Project Office, a public-private partnership that is developing four city-scale platforms for advanced wireless research in U.S. His research is supported mostly by the U.S. Federal Agencies, including the National Science Foundation, the Air Force Research Laboratory, the Office of Naval Research, and the Army Research Laboratory. His research focuses on modeling, optimization, and experimental evaluation of wireless networked systems, with applications to 5G Networks and Internet of Things, software-defined networking, and body area networks. He is a fellow of the IEEE and a Senior Member of the ACM. He was a recipient of the National Science Foundation CAREER Award. He is the Technical Program Committee Chair of the IEEE INFOCOM 2018. He is an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON BIOLOGICAL, MOLECULAR, AND MULTI-SCALE COMMUNICATIONS, *Computer Networks*, and *Smart Health*.

Sergio Palazzo received the degree in electrical engineering from the University of Catania, Catania, Italy, in 1977. Since 1987, he has been with the University of Catania, where is currently a Professor of telecommunications networks. In 1994, he spent the summer at the International Computer Science Institute, Berkeley, as a Senior Visitor. In 2003, he was with the University of Canterbury, Christchurch, New Zealand. His current research interests are in modeling, optimization, and control of wireless networks, with applications to cognitive and cooperative networking, SDN, and sensor networks. He was a recipient of the Visiting Erskine Fellowship. He has been serving on the Technical Program Committee for INFOCOM since 1992. He has been the General Chair of some ACM conferences (MobiHoc 2006 and MobiOpp 2010). He has also been the TPC Co-Chair of some other conferences, including IFIP Networking 2011, IWCMC 2013, and European Wireless 2014. He also served on the Editorial Board of several journals, including the IEEE/ACM TRANSACTIONS ON NETWORKING, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the *IEEE Wireless Communications Magazine*, *Computer Networks*, *Ad Hoc Networks*, and *Wireless Communications and Mobile Computing*.

Salvatore D'Oro (S'12–M'17) received the Ph.D. degree from the University of Catania in 2015. He is currently an Associate Research Scientist with Northeastern University. In 2015, 2016, and 2017, he organized the 1st, 2nd, and 3rd Workshops on COmpetitive and COoperative Approaches for 5G networks (COCOA). He also served on the Technical Program Committee (TPC) of the CoCoNet8 Workshop at IEEE ICC 2016, the IEEE Conference on Standards for Communications and Networking in 2018, and Med-Hoc-Net 2018. In 2013, he served on the TPC of the 20th European Wireless Conference in 2014. His research interests include game-theory, optimization, learning, and their applications to telecommunication networks. He is a member of the IEEE.