Optimal User-Cell Association for Massive MIMO Wireless Networks

Dilip Bethanabhotla\textsuperscript{1}, Ozgun Y. Bursalioglu\textsuperscript{2}, Haralabos C. Papadopoulos\textsuperscript{2}, and Giuseppe Caire\textsuperscript{1}

Abstract

The use of a very large number of antennas at each base station site (referred to as “Massive MIMO”) is one of the most promising approaches to cope with the predicted wireless data traffic explosion. In combination with Time Division Duplex and with simple per-cell processing, it achieves large throughput per cell, low latency, and attractive power efficiency performance. Following the current wireless technology trend of moving to higher frequency bands and denser small cell deployments, a large number of antennas can be implemented within a small form factor even in small cell base stations. In a heterogeneous network formed by large (macro) and small cell base stations, a key system optimization problem consists of “load balancing”, that is, associating users to base stations in order to avoid congested hot-spots and/or under-utilized infrastructure.

In this paper, we consider the user-base station association problem for a massive MIMO heterogeneous network. We formulate the problem as a network utility maximization, and provide a centralized solution in terms of the fraction of transmission resources (time-frequency slots) over which each user is served by a given base station. Furthermore, we show that such a solution is physically realizable, in the sense that there exists a sequence of integer scheduling configurations realizing (by time-sharing) the optimal fractions. While this solution is optimal, it requires centralized computation. Then, we also consider decentralized user-centric schemes, formulated as non-cooperative games where each user makes individual selfish association decisions based only on its local information. We identify a class of schemes such that their Nash equilibrium is very close to the global centralized optimum. Hence, these user-centric algorithms are attractive not only for their simplicity and fully decentralized implementation, but also because they operate near the system “social” optimum.

Index Terms

Massive MIMO, Small Cells, Scheduling, Load Balancing, User-Base Station Association, Game Theory, Nash Equilibrium.

\textsuperscript{1} Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA. Email: bethanab, caire@usc.edu.

\textsuperscript{2} Wireless Systems Project, Docomo Innovations Inc, Palo Alto, CA 94304, USA. Email: obursalioglu, hpapadopoulos@docomoinnovations.com
I. INTRODUCTION

With the proliferation of mobile devices and services, industry predicts that the mobile/wireless data traffic is going to increase by two to three orders of magnitude within a decade [1]. Although the definition of the next generation of systems and standards is at its initial phase, it is widely agreed that the next generation of wireless networks, generally referred to as “5G”, will involve a combination of multiuser MIMO technology, cell densification, and heterogeneous architectures based on nested tiers of smaller and smaller cells operating at higher and higher frequencies, in order to target traffic hot-spots and eventually meeting the traffic demands [2]. These trends have motivated the recent surge of research on massive and dense deployment of base station antennas, both in the form of “Massive MIMO” schemes, with hundreds of antennas at each cell site [3]–[5], and in the form of multi-tier networks of densely deployed small-cells [6], [7].

Massive MIMO promises large increases in spectral efficiency by transmitting independent data streams simultaneously to multiple users sharing the same transmission resource (time-frequency slot). The massive MIMO regime [3]–[5] distinguishes itself from classical multiuser MIMO [8], [9] by the fact that the number of served users is much less than the (very large) number of base station antennas. Operating in TDD mode, massive MIMO can provide very large spectral efficiencies with low latency, simple per-cell processing, and very attractive power efficiency due to the large array gain [3]. Thanks to the higher and higher carrier frequencies [10], it is possible to implement massive MIMO schemes even in small cell base stations within a reasonable form factor. Hence, it is envisaged that massive MIMO will not just be applied to large tower-mounted base stations, but also used in conjunction with small cells.

a) The load balancing problem (overview of the state of the art): In such a heterogeneous cell deployment, a key system optimization problem consists of “load balancing” or, more generally, the optimal association of users to base stations. The traditional minimum-distance association leading to a Voronoi tessellation of the coverage area (e.g., think of the classical textbook case of hexagonal cells) may be dramatically inefficient in the presence of a non-uniform cell deployment with heterogeneous base stations characterized by different transmit power and number of antennas, and/or a non-uniform user spatial distribution. In conventional technologies, user-base station association is decided on the basis of the maximum received beacon signal strength (usually denoted as Received Signal Strength

\footnote{For example, at 30GHz (\(\lambda = 1cm\)), it is possible to fit a 100-antenna 2D array on a \(10 \times 10cm\) surface with one-\(\lambda\) antenna separation.
Indicator (RSSI) and indicated by the “five bars” display on every mobile handset \[11\]. In general, such an approach can be arbitrarily suboptimal in heterogeneous scenarios. “Biasing” is a commonly proposed heuristic method for heterogeneous networks, where the RSSI is artificially scaled by a “bias” term that depends on the base station tier, to inherently steer the user association towards lower-tier base stations and therefore “off-load” the macro-cell. Many other options exist, especially, for single antenna systems (see \[12\] and the references therein). A more systematic approach to user-base station association as a combinatorial optimization problem was taken in \[12\]–\[15\]. Alternatively, a game theoretic user-centric approach is considered in \[16\], where users make randomized association decisions based only on local information, trying to selfishly optimize their own objective function. A common theme in the works \[12\]–\[15\] is to first formulate a combinatorial optimization problem where a network-wide objective function is maximized under the constraint that each user must be associated permanently to a single BS. Then, the combinatorial optimization problem is relaxed to a convex program which is efficiently solved yielding fractional solutions. These fractional solutions are then rounded back to integer feasible configurations, enforcing each user to uniquely associate with a single BS. For instance, in \[12\] the authors formulate a relaxed convex program for the special case of proportional fairness objective function (with some approximation), and they provide a decentralized algorithm which is essentially the iterative solver for their convex program, yielding a fractional solution. This solution is then rounded back to a feasible configuration of the original combinatorial optimization problem, enforcing unique association. The work in \[12\], however, does not consider the possibility of implementing fractional solutions and in fact it does not generally arrive at a fractional solution which is physically realizable (this concept is made precise in the sequel). On a separate thread, \[16\] proposes a user-centric game-theoretic approach to the association problem, which is completely decentralized. The associated randomized algorithm (which can be turned into an on-line protocol) is shown to converge to a Nash equilibrium under certain conditions on the per-user utility function. The Pareto efficiency of the Nash equilibria is studied, but it is not a priori clear whether such operating points are close to any well-defined global system optimality (social welfare).

\[b\)] Contributions of this work: In this paper we focus on the problem of optimal user-base station association for a heterogeneous wireless network with massive MIMO base stations. We formulate a Network Utility Maximization (NUM) convex problem in terms of a global network utility function reflecting some desired notion of fairness, where optimization is with respect to the fractions of transmission resources over which each user is associated to a given base station. Furthermore, we show that the solution to the convex problem is physically realizable in the following sense: there exists a feasible schedule consisting of a sequence of integer scheduling configurations such that, by time-sharing these
configurations over this schedule, the achieved time-averaged user rates can be made arbitrarily close to the rates obtained by the optimal fractional solution. While our solution is optimal, its implementation as an on-line protocol requires centralized computation and coordination across the base stations. This may be undesirable for a practical implementation. We also consider fully decentralized user-centric schemes similar to [16]. In particular, we prove conditions for which the Nash equilibrium of these user-centric schemes is very close to the global centralized optimum. Hence, this class of algorithms are not only attractive for their simplicity and fully decentralized implementation, but also because they provably operate near the system “social” optimum.

Our work differs from the optimization formulation of [12] since in our case we can prove that a fractional optimal solution is in fact physically realizable. Despite the practical implementation difficulty, this can serve as a benchmark for any decentralized low-complexity association scheme. Furthermore, we extend the user-centric scheme of [16] to a more general class of user-centric utility functions reflecting a desired notion of local (per-cell) fairness, and we show the non-trivial fact that such class of schemes operate near the system social optimum of the corresponding network-wide utility function.

The rest of this paper is organized as follows. In Section III we describe the system model and the peak rates that users get from a massive MIMO base station. In Section III we formulate the load balancing problem as a combinatorial Network Utility Maximization problem and show that it admits a convex relaxation yielding a fractional solution which, as shown in the appendix, is physically realizable. We then provide a centralized algorithm based on the dual subgradient method in Section IV to solve the convex relaxation. In Section V we consider the proposed class of user-centric schemes and prove conditions for their global near-optimality. Finally, in Section VI we verify the results through simulations in a realistic network scenario.

II. SYSTEM MODEL

We consider a system formed by $J$ base stations (BSs) serving $K$ single antenna users, distributed over a given coverage area. We use $j \in \mathcal{J} = \{1, 2, \ldots, J\}$ and $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ to index BSs and users respectively. Each BS schedules transmissions over contiguous time-frequency slots, referred to as resource blocks (RBs), each comprising a block of OFDM subcarriers and symbols.\footnote{For example, in LTE [17], resource blocks are 7 OFDM symbols long (corresponding to a duration of 0.5 ms), and 12 subcarriers wide (corresponding to a bandwidth of $12 \times 15$kHz $= 180$kHz).} We model the wireless channel as block-fading including large-scale and small-scale effects. The large-scale channel
coefficients are a function of the BS-user distance and shadowing, and are assumed to be constant across a large number of RBs. The small-scale effects are modeled as Rayleigh fading coefficients that remain constant within each RB, and evolve across different RBs according to a wide-sense stationary process with given time-frequency correlation \[17\]. We let \( M_j \) denote the number of antennas at BS \( j \), and \( S_j \) denote the number of downlink data streams that BS \( j \) can transmit on any given RB. We assume TDD operation with reciprocity-based channel state estimation \[3\], \[5\]. Hence, every BS antenna in the vicinity of user \( k \) can estimate its downlink channel coefficient to user \( k \) from the uplink pilot transmitted by user \( k \). In this way, every user \( k \) can train an arbitrarily large number of antennas in its vicinity. This enables the training of large antenna arrays (e.g., \( M_j \gg 1 \)) with training overhead proportional to \( S_j \). \[3\].

Within any RB, each BS serves in parallel a subset of the users associated to it using multiuser MIMO precoding. For simplicity, we focus on the case where each BS serves a fixed-size set of users with linear zero-forcing beamforming (LZFBF) \[18\], using equal power per data stream. We hasten to say that the requirement of equal-power allocation can be relaxed and the optimal power allocation in the massive MIMO regime (1 \( \ll \) \( S_j \ll M_j \)) can be addressed using the results in \[18\], at the cost of some analytical complication in the rate formulas. Since here we focus on the structure of the problem solution and this complication is inessential, we shall restrict to uniform power allocation across the downlink data streams. Regarding the number of downlink data streams per BS, it is observed in \[3\], \[5\] that, in TDD multiuser MIMO systems, the number of users that can be served by a BS using multiuser MIMO depends on the dimension of the uplink pilot field, which in turn determines the number of user channel vectors that can be estimated and for which the downlink precoder can be calculated. The length of the pilot field, in turn, depends on the channel coherence time, since uplink channel estimation and downlink precoding must take place in the same channel coherence interval in order to achieve a small channel estimation error (this has been demonstrated including TDD uplink-downlink reciprocity calibration in \[19\], \[20\]).

Although a lot of attention has been devoted in the literature to the problem of “pilot contamination”, i.e., the bias in the channel estimation due to users sharing the same pilot symbol sequence in different cells, it is known that with LZFBF precoding and when the ratio \( S_j/M_j \) is not vanishingly small, the pilot contamination interference power contributes for a relatively small amount to the whole inter-cell interference \[4\], \[5\]. Hence, in the rest of this paper we shall neglect imperfect estimation effects, and
assume perfect downlink channel state knowledge at the BSs.

A. Modeling the Massive MIMO User Rates

Following [5], we refer to the massive MIMO regime as the case where $1 \ll S_j \ll M_j$. In this regime, the user rates tend to concentrate on deterministic limits, i.e., the effect of the small-scale fading disappears, due to the well-known channel hardening effect [21]. Such limits can be calculated using asymptotic random matrix theory in a great deal of generality with respect to the statistics of the channel vectors [22], [23]. In this section, we borrow such results for the well-known case of LZFBF precoding and i.i.d. Rayleigh fading. This yields very simple, yet accurate, rate formulas that shall be used in the system optimization problem developed in the next section.

Letting, $g_{jk}$ denote the large-scale fading gain between BS $j$ and user $k$, and $P_j$ denote the transmit power of BS $j$, the rate $R_{k,j}$ that can be reliably supported by transmitting from BS $j$ to user $k$ can be approximated by:

$$R_{k,j} = \log \left( 1 + \frac{M_j - S_j + 1}{S_j} \frac{P_j g_{jk}}{1 + \sum_{j' \neq j} P_{j'} g_{j'k}} \right).$$

This approximation holds exactly (see for example [18]) in the limit of $M_j, S_j \rightarrow \infty$ with fixed ratio $S_j/M_j = \nu_j$, where $\nu_j$ denotes the dimensional load of BS $j$, i.e., the number of transmitted data streams over the number of BS antennas.

The following observations are in order: 1) Using the deterministic approximation (1) instead of the actual user rates, which are a function of the (random) $M_j \times S_j$ channel matrix between the BS and the users’ antennas) has the advantage of providing a clean analytical and very accurate handle for the system optimization; 2) It is known that the (a.s.) convergence to the deterministic approximation is very quick with respect to the problem dimension $M_j$. In particular, a central limit theorem can be proved such that for large but finite $M_j$ the actual rate can be written as $R_{k,j} + \chi_{k,j}$, where $\chi_{k,j}$ is a Gaussian “fluctuation” with mean zero and variance $O(1/M_j^2)$) (see for example [23] and references therein). It follows that the rate expression (1) is very accurate for practical values of $M_j$ and $S_j$ [18]; 3) The convergence of the massive MIMO user rates to the deterministic limit (1) indicates that, in practice, it is not necessary to adapt the transmission rate on a per-RB basis. Instead, it is sufficient to encode information at rate

---

3With simple conjugate beamforming instead of LZFBF and in the regime of finite $S_j$ and $M_j \rightarrow \infty$, as considered in [3], pilot contamination is the only source of residual inter-cell interference. However, it has been shown in [5] that this regime is reached for an impractical number of antennas (around $10^4$). Furthermore, the effect of pilot contamination with LZFBF is even less important.
for some back-off margin $\epsilon > 0$, and interleave the physical layer codewords over a sequence of RBs. Assuming a powerful channel code (e.g., a Turbo Code or an LDPC code as those specified in modern wireless standards [24]), a sufficiently small block error rate can be achieved. We conclude that, for the sake of analytical tractability, it makes sense to take (1) as the achievable user coding rate for user $k$ served by BS $j$.

Next, we consider the user throughput, i.e., the long-term average rate achieved by a user over a sequence of RBs. To this regard, we notice that the number of users associated to a given BS $j$ may be larger than $S_j$ (which is fixed a priori according to the pilot dimension constraint as discussed before). This means that users are not “active” in each RB. Instead, a scheduling algorithm decides which subset of size $S_j$ users is active over any RB, according to some criterion. We denote by $\alpha_{k,j}$ the scheduling activity fraction of user $k$ at BS $j$, i.e., the fraction of RBs over which user $k$ is in the scheduled active subset of BS $j$. It follows that, by definition, the throughput that user $k$ receives from BS $j$ is given by $\alpha_{k,j}R_{k,j}$. In the following, we shall refer to the rates $R_{k,j}$ given by (1) as “peak” rates, and to the product $\alpha_{k,j}R_{k,j}$ as (long-term average) throughput rates.

### III. Load Balancing Problem Formulation

We wish to find the optimal association of users to BSs such that an overall network utility function $U(r)$ of the user throughputs vector $r = (r_1, \ldots, r_k)$ is maximized. We shall choose the network utility function in order to achieve a desired balance between network-wide overall performance and user fairness. For example, it is well-known that maximizing the sum throughput in a cellular system may lead to solutions where only the users close to the cell centers are served, while the users at the cell edge are left starving (see for example [5], [18]). Such a solution is clearly not acceptable. In contrast, desirable network utility functions $U(r)$ are concave and componentwise monotonically increasing, such that larger user throughputs yield larger utility, but the shape of the concave function imposes some desired notion of fairness. For example, Proportional Fairness (PF) (equal air time) is imposed by choosing the utility function $U(r) = \sum_k \log r_k$. Instead, Hard Fairness (HF) (maximizing the minimum rate over all users) is imposed by choosing the utility function $U(r) = \min_k r_k$. These are special cases of the class of utility functions defined by [25]

$$U(r) = \sum_k \phi_\gamma(r_k),$$  \hspace{1cm} (2)

where $\gamma \geq 0$ is a parameter that determines the level of fairness and

$$\phi_\gamma(x) = \begin{cases} 
\log x & \text{for } \gamma = 1 \\
\frac{x^{1-\gamma}}{1-\gamma} & \text{for } \gamma \neq 1 
\end{cases}$$  \hspace{1cm} (3)
This class includes PF ($\gamma = 1$), HF ($\gamma \to \infty$) and the unfair sum-throughput case ($\gamma = 0$).

Let $\mathcal{J}_k \subseteq \mathcal{J}$ denote the set of BSs which can potentially serve user $k$; the elements in this set are, e.g., the BSs $j$ with sufficiently high peak rate $R_{k,j}$, and/or belonging to some restricted set (in the case of restricted access due to other network constraints). We assume single-cell processing, i.e., over any RB a user $k$ is served by at most one $\text{BS } j \in \mathcal{J}_k$. It is important to notice that, though user $k$ is served by a single $\text{BS in } \mathcal{J}_k$ in a single RB due to single-cell processing, it may be served by different BSs in $\mathcal{J}_k$ on different RBs. Consequently, a user $k$ can be associated fractionally to multiple BSs in $\mathcal{J}_k$, with $\alpha_{k,j}$ denoting the fraction of RBs over which user $k$ is served by BS $j$ (activity fraction), as already defined. We express this notion formally through the following definitions:

**Definition 1. Association:** A user $k$ is said to be associated to the set of BSs $\mathcal{J}^*_k \subseteq \mathcal{J}_k$ if $\alpha_{k,j} > 0$ for all $j \in \mathcal{J}^*_k$ and $\alpha_{k,j} = 0$ for all $j \in \mathcal{J}_k \setminus \mathcal{J}^*_k$. If $\alpha_{k,j} = 0$ for all $j \in \mathcal{J}_k$, we say that user $k$ is not associated, i.e., $\mathcal{J}^*_k = \emptyset$.

**Definition 2. Unique Association:** A user $k$ is said to be uniquely associated to BS $j^*$ if $\alpha_{k,j} > 0$ for $j = j^*$ and $\alpha_{k,j} = 0$ for every $j \neq j^*$ (i.e., in this case $|\mathcal{J}^*_k| = 1$).

Notice that, even though user $k$ is uniquely associated to BS $j^*$, it is not necessarily served by BS $j^*$ on all RBs. For example, if $\alpha_{k,j^*} = 0.5$, then BS $j^*$ serves user $k$ only on 50% of the RBs.

**Definition 3. Fractional Association:** A user $k$ is said to be fractionally associated to the set of BSs $\mathcal{J}^*_k \subseteq \mathcal{J}_k$ if $|\mathcal{J}^*_k| > 1$.

---

4It is possible and often beneficial to serve user $k$ by jointly precoding antennas across multiple BSs in $\mathcal{J}_k$. Such coordinated multi-point (CoMP) schemes have been widely studied in the context of massive MIMO (e.g., see [5] and references therein). Nevertheless, here we focus on single-cell processing which is more desirable from a practical viewpoint and more in line with the massive MIMO philosophy [3].
At this point, the Network Utility Maximization (NUM) problem at hand can be expressed by:

\[
\text{maximize } U(r) \quad (4a)
\]

subject to

\[
\begin{align*}
\sum_{j \in J_k} \alpha_{k,j} r_{k,j} & \leq \sum_{j \in J_k} \alpha_{k,j} R_{k,j}, \ \forall \ k \in K \\
\sum_{k \in K} \alpha_{k,j} & \leq S_j, \ \forall \ j \in J \\
\sum_{j \in J} \alpha_{k,j} & \leq 1, \ \forall \ k \in K \\
r_k & \geq 0, \ \alpha_{k,j} \geq 0, \ \forall \ k \in K, \ j \in J,
\end{align*}
\] 

(4b)–(4e)

where \( R_{k,j} \) are the peak rates given by (1) and where the optimization is with respect to \( r \) and \( \alpha = \{\alpha_{k,j}\} \).

A brief explanation of the constraints (4b)–(4e) is in order:

- The constraint (4b) follows simply from the definition of the aggregate throughput of user \( k \) over all its serving BSs \( j \in J_k \) and from the fact that \( U(\cdot) \) is componentwise increasing, such that the optimum is always achieved when (4b) is satisfied with equality for all \( k \).
- The constraints in (4c) reflect the fact that the sum activities of all the users being served by any given BS \( j \) cannot exceed the number of simultaneous downlink data streams \( S_j \) (transmit spatial degrees of freedom constraint).
- The constraint (4d) simply reflects the fact that each user’s total activity fraction (over all BSs) cannot be more than one (achieving the bound with equality means that a user is served in every resource block).

We now define a few terms which will be useful in the sequel.

**Definition 4. Feasible Association Configuration:** Any set \( \{\alpha_{k,j}\} \) which satisfies the constraints (4c)-(4e) is termed a feasible association configuration.

\[ \text{Definition 5. Unique Association Configuration:} \text{ A feasible association configuration is said to be a unique association configuration if every user } k \in K \text{ is uniquely associated to some BS, i.e., if } |J_k^*| = 1 \ \forall \ k \in K. \]

\[ \text{Definition 6. Integer Scheduling Configuration:} \text{ A feasible association configuration is said to be an integer scheduling configuration if } \alpha_{k,j} \in \{0, 1\} \text{ for all pairs } (k, j). \]

Note that the problem formulation (4) is different from previous formulations [12], [14], [15] whereby each user is constrained to associate permanently to a single BS and where, as a consequence, there
an additional set of constraints reflecting the unique association constraint for each user, yielding a combinatorial optimization problem. Although unique association configurations are practically attractive (as discussed in the sequel), our NUM problem formulation in (4) has a number of key desirable features:

(i) Problem (4) is a convex problem. We shall develop in Section IV an efficient method for its solution that also sheds light on the properties of the optimal solution.

(ii) Theorem 1, given in Appendix A, establishes that any feasible association configuration is physically realizable, i.e., given a feasible association configuration \( \{\alpha_{k,j}\} \), we can produce a deterministic sequence of integer scheduling configurations such that the long-term averaged user activity fractions over such a sequence are arbitrarily close to the given feasible association configuration \( \{\alpha_{k,j}\} \).

(iii) Since (4) is a relaxation of the unique association combinatorial optimization problem, and since its solution can actually be implemented, the utility function value achieved by its (feasible) solution upper-bounds the value of the solution of any fixed unique association configuration. This means that the solution of (4) yields a feasible benchmark with respect to which any practical (e.g., decentralized, efficiently computable) user-BS association and load balancing scheme should be compared.

Remark 1. An immediate consequence of the network utility function in (2) is that, for \( \gamma \geq 1 \), the solution of (4) must associate all users, i.e., for all \( k \in K \) it must be \( |J_k^*| \geq 1 \). Otherwise, some user would have zero throughput, yielding \( U(r) = -\infty \). The feature that all users in the system are served with non-zero throughput reflects the notion of fairness built into the NUM problem and it is very desirable in practice. Therefore, from now on, we shall restrict to \( \gamma \geq 1 \).

Remark 2. The problem (4) can be extended so that each BS \( j \) can optimize also over the discrete choice of the number of downlink streams, i.e., over \( S_j \in \{1, 2, \ldots, S_j^{\text{max}}\} \), for some \( S_j^{\text{max}} \leq M_j \). This extension, however, yields a combinatorial problem that can be decomposed into \( \prod_{j=1}^J S_j^{\text{max}} \) subproblems, each of which takes the form (4) for a fixed choice of \( \{S_j\} \). In a practical scenario with tens of BSs each of which can serve a maximum of tens of downlink data streams, the number of such subproblems becomes quickly intractable (of the order of \( 10^{10} \)). We believe that, in practice, optimizing over the number of downlink data streams is quite irrelevant, since in general we have more users than downlink streams (i.e., there is no point in decreasing \( S_j \) below its maximum) and the maximum value of \( S_j \) is fixed by other technology constraints (e.g., the uplink pilot overhead, as discussed before).

A related class of convex relaxations of the unique association problem is also considered in [12] for the single-input single output (SISO) case. Although the relaxations in [12] provide upper bounds
to the unique association solution, the activity fractions yielding these upper bounds are not necessarily physically realizable. In fact, the feasible region of the relaxed convex program in [12] may contain certain fractional associations which cannot be implemented using a scheduling policy consisting of a sequence of integer scheduling configurations.

IV. CENTRALIZED SOLUTION

In order to solve the convex program (4), general purpose numerical solvers like CVX can be used [26]. However, here we focus on solving (4) by using a dual subgradient method, which is both much more efficient than general purpose solvers, and also sheds light on the properties of the solution. We first formulate the dual program of (4) for a general utility function \( U(\cdot) \). We then specialize to the class defined in (2) and develop a centralized algorithm to solve the dual program. The optimal solution to the dual program is then used to solve for the primal variables \( \alpha \).

We form the Lagrangian function for the primal problem (4) by introducing the dual variables/prices (we use the terms dual variables and prices interchangeably) \( \beta = \{ \beta_k \} \) for the constraint (45), \( p = \{ p_j \} \) for the constraint (46) and \( \lambda = \{ \lambda_k \} \) for the constraint (44). Then, the Lagrangian function takes on the form:

\[
L(\alpha, r, \beta, p, \lambda) = U(r) - \sum_k \beta_k (r_k - \sum_j \alpha_{k,j} R_{k,j}) - \sum_j p_j \left( \sum_k \alpha_{k,j} - S_j \right) - \sum_k \lambda_k \left( \sum_j \alpha_{k,j} - 1 \right)
\]

\[= U(r) - \sum_k \beta_k r_k + \sum_j S_j p_j + \sum_k \lambda_k + \sum_{(k,j)} \alpha_{k,j} (\beta_k R_{k,j} - p_j - \lambda_k).
\]

(5)

The value \( G(\beta, p, \lambda) \), for any set of non-negative dual variables, provides an upper bound on the optimal primal objective value. The dual program finds the tightest of such upper bounds by minimizing \( G(\beta, p, \lambda) \) over the feasible set of dual variables [27], i.e., it is given by:

\[
\text{minimize } G(\beta, p, \lambda) \quad \text{subject to } \beta, p, \lambda \geq 0.
\]

(6)
From (6), we observe that \( G(\beta, p, \lambda) = \infty \) if \( \beta_k R_{k,j} - p_j - \lambda_k > 0 \) for some \((k, j)\). In the case when \( \beta_k R_{k,j} - p_j - \lambda_k \leq 0 \) for all \((k, j)\), it is easy to see that \( L(\alpha, r, \beta, p, \lambda) \) is maximized when \( \alpha \) is chosen such that the term \( \sum_{(k, j)} \alpha_{k,j} (\beta_k R_{k,j} - p_j - \lambda_k) \) in (6) vanishes. From these observations, we find the dual program equivalent form:

\[
\text{minimize } \max_{r \geq 0} \left\{ U(r) - \sum_k \beta_k r_k \right\} + \sum_j S_j p_j + \sum_k \lambda_k
\]

subject to \( \beta_k R_{k,j} \leq p_j + \lambda_k \forall (k, j) \)

\( \beta, p, \lambda \geq 0. \) (8c)

We now particularize to the class of network utility functions in (2). Thanks to the additive form of the network utility function, the maximization with respect to \( r \) in (8a) decomposes into the sum (over \( k \in K \)) of individual maximizations of the terms

\[
\phi_{\gamma}(r_k) - \beta_k r_k, \quad k \in K.
\]

Setting the derivative with respect to \( r_k \) to zero, it is immediate to show that the maximum is achieved for

\[
r_k = \beta_k^{-\rho},
\]

where, for future convenience, we define \( \rho = 1/\gamma \). The corresponding (maximum) value is \( \frac{1}{\rho-1} \beta_k^{1-\rho} \) for \( \gamma \neq 1 \), and \(-\log \beta_k - 1\) for \( \gamma = 1 \). We first consider in detail the case \( \gamma \neq 1 \). Using (9) into (8a), we obtain the dual program in the form:

\[
\text{minimize } \sum_j S_j p_j + \sum_k \lambda_k + \frac{1}{\rho-1} \sum_k \beta_k^{1-\rho}
\]

subject to \( \beta_k \leq \frac{p_j + \lambda_k}{R_{k,j}} \forall (k, j) \)

\( \beta, p, \lambda \geq 0. \) (10b)

The minimization over \( \beta \) is immediate, and yields

\[
\beta_k = \min_j \left\{ \frac{p_j + \lambda_k}{R_{k,j}} \right\}.
\]

(11)

Replacing, we obtain:

\[
\text{minimize } \tilde{G}(p, \lambda) = \sum_j S_j p_j + \sum_k \lambda_k + \frac{1}{\rho-1} \sum_k \left( \min_j \left\{ \frac{p_j + \lambda_k}{R_{k,j}} \right\} \right)^{1-\rho}
\]

subject to \( p, \lambda \geq 0\). (12a)
We next describe a convergent subgradient algorithm to approximate arbitrarily closely the solution to (12). We let \((p^{(i)}, \lambda^{(i)})\) denote the value of the dual variables at the \(i\)-th subgradient iteration. The \((i+1)\)-th iterate is given as the \(i\)-th iterate minus an appropriately scaled adjustment along the subgradient chosen at the current iteration \[28\]. The subgradient algorithm comprises the following steps:

1) Initialize \((p, \lambda)\) to some arbitrary positive values \((p^{(0)}, \lambda^{(0)})\) and let \(i = 0\). Also choose the number of iterations \(i_{\text{max}}\) and the step sequence \(s^{(i)} = \frac{a}{b^{i+1}}\) with appropriately chosen constants \(a > 0\), \(b > 0\).

2) Choose the subgradient for the \(i\)-th iteration, based on the objective function in (12a) evaluated in the neighborhood of \((p^{(i)}, \lambda^{(i)})\). In particular, let:
\[
\hat{R}^{(i)}(p, \lambda) = \sum_j \alpha_j p_j + \sum_k \lambda_k + F^{(i)}(p, \lambda)
\]
where
\[
F^{(i)}(p, \lambda) = \frac{1}{\rho - 1} \sum_k \left( \frac{p_j + \lambda_k}{R_{k,j}^{(i)}} \right)^{1-\rho} = \frac{1}{\rho - 1} \sum_j \sum_{k \in K^{(i)}_j} \left( \frac{p_j + \lambda_k}{R_{k,j}^{(i)}} \right)^{1-\rho}.
\]

3) Taking derivatives of \(\hat{G}^{(i)}(p, \lambda)\) with respect to \(p_j\) and \(\lambda_k\), respectively, and the non-negativity constraint of \(p_j\) and \(\lambda_k\), the corresponding subgradient iteration is given by
\[
p_j^{(i+1)} = \left[ p_j^{(i)} + s^{(i)} \left( \sum_{k \in K^{(i)}_j} \frac{R_{k,j}^{(i-1)}}{p_j^{(i)} + \lambda_k^{(i)}} - S_j \right) \right]^+
\]
\[
\lambda_k^{(i+1)} = \left[ \lambda_k^{(i)} + s^{(i)} \left( \frac{R_{k,j}^{(i-1)}}{p_j^{(i)} + \lambda_k^{(i)}} - 1 \right) \right]^+
\]

4) If \(i < i_{\text{max}}\) increment \(i\) by 1 and go to step 2, else stop.

It can easily be verified that the formulas (13) also provide the corresponding subgradient algorithm iteration updates in the case \(\gamma = 1\).

5Alternatively we could choose a stopping criterion for the algorithm.

6In case multiple \(j\) indices maximize \(R_{k,j}/(p_j^{(i)} + \lambda_k^{(i)})\) any one of these can be used.
In the following, we denote by \((p^*, \lambda^*)\) the dual variable values after a sufficiently large number of dual subgradient iterations. Then, we need to solve for the corresponding primal variables (activity fractions) of (4), thus obtaining the optimal association configuration and the corresponding optimal user throughputs. Before describing techniques for this purpose in Section IV-B we first discuss the KKT conditions of (4).

A. KKT conditions

The convex program (4) is given in canonical form with linear inequality constraints (4b)–(4e). Therefore the Slater condition reduces to feasibility. This implies that strong duality holds and the KKT conditions including the feasibility and the complementary slackness conditions are both necessary and sufficient for optimality. Noticing that all variables are non-negative (for a classical argument, see [29, Th. 4.4.1]), by taking the partial derivatives of \(L(\alpha, r, \beta, p, \lambda)\) with respect to \(r_k\) and \(\alpha_{k,j}\) we obtain necessary and sufficient conditions for optimality in the form

\[
\frac{\partial L}{\partial r_k} = \phi'_\gamma(r_k) - \beta_k \leq 0 \\
\frac{\partial L}{\partial \alpha_{k,j}} = \beta_k R_{k,j} - p_j - \lambda_k \leq 0
\]

where inequalities (14)–(15) must hold with strict equality for the strictly positive components \(r_k, \alpha_{k,j}\), respectively, at the optimal points. In addition, the complementary slackness conditions are given in the form

\[
\sum_{k \in K} \alpha_{k,j} - S_j \leq 0 \\
\sum_{j \in J} \alpha_{k,j} - 1 \leq 0
\]

where the inequalities (16)–(17) must hold with strict equality for the strictly positive components \(p_j, \lambda_k\), respectively, at the optimal points. Let \((\alpha^\circ, r^\circ, \beta^\circ, p^\circ, \lambda^\circ)\) denote an optimal point, i.e., a set of values for \((\alpha, r, \beta, p, \lambda)\) that achieves the min (w.r.t. the dual variables) of the max (w.r.t. the primal variables) of (6), and recall that we assumed \(\gamma \geq 1\), implying that \(r_k^\circ = 0\) for some \(k\) cannot be an optimal point (see Remark 1). This implies that (14) must hold with equality at \((\alpha, r, \beta, p, \lambda) = (\alpha^\circ, r^\circ, \beta^\circ, p^\circ, \lambda^\circ)\). Using the expression \(\phi'_\gamma(x) = x^{-\gamma}\), the definition \(r_k = \sum_{j \in J_k} \alpha_{k,j} R_{k,j}\) and substituting \(\beta_k = \phi'_\gamma(r_k)\) in
\[(15), (14)-(15)\] reduce to
\[
\sum_{j \in J_k} \alpha_{k,j} R_{k,j'} = \left( \frac{R_{k,j}}{p_j^o + \lambda_k^o} \right) \quad \forall \ j \in J'_k(p^o, \lambda^o) \tag{18}
\]
\[
\sum_{j' \in J_k} \alpha_{k,j'} R_{k,j'} > \left( \frac{R_{k,j}}{p_j^o + \lambda_k^o} \right) \quad \forall \ j \notin J'_k(p^o, \lambda^o), \tag{19}
\]
where the set \(J'_k(p^o, \lambda^o)\) is given by
\[
J'_k(p^o, \lambda^o) = \left\{ j : \frac{R_{k,j}}{p_j^o + \lambda_k^o} = \max_{j' \in J_k} \frac{R_{k,j'}}{p_{j'}^o + \lambda_k^o} \right\}. \tag{20}
\]
Summarizing, the consistency conditions for the optimality of the activity fractions at an optimal set of prices are as follows: for each user \(k \in K\) and the sets of BSs defined in (20), we have
\[
\begin{cases}
\alpha_{k,j} > 0 \quad \text{for some} \ j \in J'_k(p^o, \lambda^o) \\
\alpha_{k,j} = 0 \quad \forall \ j \notin J'_k(p^o, \lambda^o). \tag{21}
\end{cases}
\]
Interpreting the quantity \(\left( \frac{R_{k,j}}{p_j^o + \lambda_k^o} \right)^\rho\) as the “bang-per-buck” offered by BS \(j\) to user \(k\) (a term from economics [30]) at the prices \(p_j^o\) and \(\lambda_k^o\), \(J'_k(p^o, \lambda^o)\) is the set of BSs offering the maximum bang-per-buck to user \(k\). Thus, the conditions (18) and (21) imply that, at the optimum prices \(p^o\) and \(\lambda^o\),
- the long term average rate of every user \(k\) should be equal to the maximum bang-per-buck value offered by some BS in its neighborhood;
- every user \(k\) can have a strictly positive activity fraction to only those BSs which offer the maximum bang-per-buck value, among the BSs in \(J_k\) (from which user \(k\) is allowed to get service).

**B. Solving for the Primal Variables**

We now describe how to use the solution \((p^o, \lambda^o)\) of the dual problem (12) to solve for the primal variables. First we note that, using (9) and (11) we have
\[
r_k^* = \left( \max_j \left( \frac{R_{k,j}}{p_j^o + \lambda_k^o} \right) \right) \quad \forall k \in K. \tag{22}
\]
Hence, the optimal user throughputs are given by (22). In order to calculate the optimal association configuration \(\alpha^*\), having obtained the optimal dual variables \((p^o, \lambda^o)\), we can solve the KKT conditions.
In particular, we can choose any feasible set of activity fractions satisfying

\[
\alpha_{k,j}^* = 0 \quad \forall \ j \notin \mathcal{J}_k^*(\mathbf{p}^*, \lambda^*)
\]  

(23a)

\[
\sum_{j \in \mathcal{J}_k^*(\mathbf{p}^*, \lambda^*)} \alpha_{k,j}^* R_{k,j} > \left( \frac{R_{k,j}}{p_j^* + \lambda_k^*} \right) \quad \forall \ j \notin \mathcal{J}_k^*(\mathbf{p}^*, \lambda^*)
\]  

(23b)

\[
\sum_{j \in \mathcal{J}_k^*(\mathbf{p}^*, \lambda^*)} \alpha_{k,j}^* R_{k,j} = \left( \frac{R_{k,j}}{p_j^* + \lambda_k^*} \right) \quad \forall \ j \in \mathcal{J}_k^*(\mathbf{p}^*, \lambda^*)
\]  

(23c)

\[
p_j^* \left( \sum_{k \in \mathcal{K}} \alpha_{k,j}^* - S_j \right) = 0 \quad \forall \ j \in \mathcal{J}
\]  

(23d)

\[
\lambda_k^* \left( \sum_{j \in \mathcal{J}} \alpha_{k,j}^* - 1 \right) = 0 \quad \forall \ k \in \mathcal{K}
\]  

(23e)

However, in practice, the dual subgradient algorithm yields dual variables that differ from their optimal value by some very small numerical error, due to finite machine precision and finite number of iterations. Hence, the system of KKT conditions above may not have a solution when \( \mathbf{p}^* \) and \( \lambda^* \) are numerically calculated. We therefore propose a numerically stable approach that always yields a feasible association configuration. As stated in Lemma 1 below, the proposed approach yields the exact optimal association configuration \( \alpha^* \) when the dual variables \( \mathbf{p}^* \) and \( \lambda^* \) are exactly at their optimal point. It is expected that in the presence of small perturbations around such an optimal point, the proposed approach yields a feasible association configuration which is very close to optimal, since this is the result of a well-behaved Linear Program (LP) with bounded coefficients and compact feasible set.

As noticed before, the optimal throughput values \( \mathbf{r}^* \) are given by (22). Define the ratios \( \tilde{R}_{k,j} = R_{k,j}/r_k^* \), and consider the quantities \( f_k(\alpha) = \sum_{j \in \mathcal{J}_k} \alpha_{k,j} \tilde{R}_{k,j} \), for \( k \in \mathcal{K} \). By construction, there exists an optimal feasible association configuration \( \alpha^* \) such that \( f_k(\alpha^*) = 1 \), i.e., there is an optimal point where the (linear) functions \( f_k(\cdot) \) are equal to 1, for all \( k \in \mathcal{K} \). This suggests that \( \alpha^* \) can be found as the solution of the LP:

maximize \( \theta \)  

subject to \( \theta \leq f_k(\alpha), \ \forall \ k \in \mathcal{K} \)  

(24a)

(24b)

\[
\sum_{k \in \mathcal{K}} \alpha_{k,j} \leq S_j, \ \forall \ j \in \mathcal{J}
\]  

(24c)

\[
\sum_{j \in \mathcal{J}} \alpha_{k,j} \leq 1, \ \forall \ k \in \mathcal{K}
\]  

(24d)

\[
\alpha_{k,j} \geq 0, \ \forall \ k \in \mathcal{K}, \ j \in \mathcal{J}.
\]  

(24e)

We have:
Lemma 1. If the throughputs $r_k^*$ given by (22) correspond to the exact optimal solution of the NUM problem (4), then the solution of (24) is the corresponding optimal feasible association configuration.

Proof: Intuitively, the LP (24) maximizes the minimum $f_k(\cdot)$ by maximizing a common lower bound $\theta$ subject to the feasibility of the association configuration. Let $\hat{\alpha}$ denote the solution of (24) and let $\theta_{\text{max}} = \min_k f_k(\hat{\alpha})$ denote the achieved maximum value of the common lower bound. Since, by construction, there exist a feasible configuration $\alpha^*$ for which $f_k(\alpha^*) = 1$ for all $k \in K$, there are two possible cases: 1) $\theta_{\text{max}} < 1$, or 2) $\theta_{\text{max}} \geq 1$. Case 1) is impossible, otherwise $\hat{\alpha}$ could be improved to $\alpha^*$, contradicting the assumption that $\hat{\alpha}$ is the solution of (24). Case 2) can only hold with equality. In fact, if it held with strict inequality, we would have $f_k(\hat{\alpha}) > 1$ for all $k$, implying

$$\sum_{j \in J_k} \hat{\alpha}_{k,j} R_{k,j} > \sum_{j \in J_k} \alpha^*_{k,j} R_{k,j} = r_k^*, \quad \forall k \in K.$$  

Since the network utility function $U(\cdot)$ is componentwise increasing, this means that there exists a feasible association configuration $\hat{\alpha}$ yielding better utility than the optimal $\alpha^*$, thus leading to a contradiction. It follows that it must be $\theta_{\text{max}} = 1 = f_k(\hat{\alpha})$ for all $k \in K$, implying $\alpha^* = \alpha^*$ since, by the system model setup, $R_{k,j} > 0$ for all $j \in J_k$.

Finally, it is important to note that once the optimal feasible association configuration $\alpha^*$ is obtained, we still need to find a physically realizable schedule, the existence of which is guaranteed by Theorem 1. If $\alpha^*$ is a unique association configuration, such a schedule is trivially obtained: define $K_j = \{ k \in K : j^* = j \}$ to denote the set of users uniquely associated to BS $j$. Then, it is sufficient that each BS $j$, independently of the other BSs, schedules its own users according to some randomized scheme such that the probability that user $k \in K_j$ is selected in any RB is $\alpha^*_{k,j}$. However, if $\alpha^*$ is not a unique association configuration, finding such a schedule is not an easy problem, in general, since the scheduling decisions at the different BSs are entangled by the presence of fractionally associated users and the different BSs need to avoid scheduling such users on the same RB. Motivated by this consideration, in the next section we study a class of schemes which yield a unique association configuration and can be implemented in a fully decentralized user-centric manner. Yet, somehow surprisingly, these schemes perform close to the globally optimal centralized solution.

V. DISTRIBUTED LOAD-BALANCING ALGORITHMS

In this section, we focus on user-centric load balancing algorithms where each user makes its own association decisions in a selfish way, i.e., based on its own user-centric utility function. In particular, we consider a class of such schemes where the user-centric utility function is the user throughput, $r_k$, and
each BS applies a local version of the NUM (4). Letting $\mathcal{K}_j$ denote the set of users uniquely associated to BS $j$, the service policy at each BS $j \in \mathcal{J}$ solves the following local NUM problem:

\[
\begin{align*}
\text{maximize} \quad & \sum_{k \in \mathcal{K}_j} \phi_\gamma(\alpha_{k,j}R_{k,j}) \\
\text{subject to} \quad & \sum_{k \in \mathcal{K}_j} \alpha_{k,j} \leq S_j, \\
& 0 \leq \alpha_{k,j} \leq 1, \quad \forall \ k \in \mathcal{K}_j.
\end{align*}
\] (25a)

Assuming as usual $\gamma \geq 1$, the solution of (25) is immediate (details are omitted for brevity) and yields

\[
\alpha_{k,j} = \begin{cases} 
\min \left\{ \frac{S_j R_{k,j}^{\gamma-1}}{\sum_{k' \in \mathcal{K}_j} R_{k',j}^{\gamma-1}}, 1 \right\}, & \text{for } j = j(k) \\
0, & \text{for } j \neq j(k)
\end{cases}
\] (26)

where $j(k)$ denotes the BS to which user $k$ is uniquely associated (i.e., $j(k) = j \iff k \in \mathcal{K}_j$). Since the activity fractions $\{\alpha_{k,j}\}$ are fixed by the local NUM policy (26), we shall not distinguish any longer between the partition and the induced unique association configuration. In particular, we have:

**Definition 7.** A partition $\{\mathcal{K}_j : j \in \mathcal{J}\}$ of the user set $\mathcal{K}$ is valid if it corresponds to a feasible unique association configuration, i.e., if $j(k) \in \mathcal{J}_k$ for all $k$.

For a given valid partition $\{\mathcal{K}_j\}$, the throughput of each user $k$ is given by

\[
r_k = \alpha_{k,j(k)}R_{k,j(k)} = \min \left\{ \frac{S_{j(k)} R_{k,j(k)}^{\gamma-1}}{\sum_{k' \in \mathcal{K}_{j(k)}} R_{k',j(k)}^{\gamma-1}}, R_{k,j(k)} \right\},
\] (27)

where the term $\sum_{k' \in \mathcal{K}_{j(k)}} R_{k',j(k)}^{\gamma-1}$ in the denominator of (27) is interpreted as a measure of the load at BS $j(k)$. For example, for the case of PF ($\gamma = 1$) this quantity is actually equal to $|\mathcal{K}_{j(k)}|$, the number of users uniquely associated with BS $j(k)$.

A. User-centric association games

Following [16], the user-centric association algorithms considered in this work can be formulated in the framework of non-cooperative association games with $K$ players (users) and $J$ resources (base stations). Each user can associate with only one BS. Thus, each player $k$ has a finite set of actions $\mathcal{J}_k$ to choose from. As said before, the payoff function of user $k$ is its throughput $r_k$, given by (27). This can be seen as a function that maps the action space $\mathcal{J}_1 \times \cdots \times \mathcal{J}_K$ to the non-negative real line. In order to express the dependency of the user throughputs on the association configuration we will use the notation $r_k(j)$. A unique association configuration $j = (j(1), \ldots, j(K))$ is said to be a pure Nash equilibrium if, for
all $k \in \mathcal{K}$, we have $r_k (j(1), \ldots, j(k), \ldots, j(K)) \geq r_k (j(1), \ldots, j, \ldots, j(K))$ for any $j \in \mathcal{J}_k$. In other words, no user has an incentive to change unilaterally its association while all other users stay unchanged.

Before discussing a specific user centric algorithm in terms of its on-line decentralized implementation, let’s examine the global optimality properties of the Nash equilibria of the related association game. Suppose that there exists a valid partition $\{ \mathcal{K}_j : j \in \mathcal{J} \}$ of the user set $\mathcal{K}$ with the following properties:

$$\frac{S_j R_{k,j}^{\rho-1}}{\sum_{k' \in \mathcal{K}_j} R_{k',j}^{\rho-1}} \leq 1, \quad \forall \ j \in \mathcal{J} \quad \text{and} \quad k \in \mathcal{K}_j \quad (28a)$$

$$\frac{S_j(k) R_{k,j}^{\rho}(k)}{\sum_{k' \in \mathcal{K}_j(k)} R_{k',j}^{\rho-1}(k)} > \frac{S_l R_{k,l}^{\rho}}{\sum_{k' \in \mathcal{K}_l} R_{k',l}^{\rho-1}}, \quad \forall \ k \in \mathcal{K}, \quad \text{and} \quad \ell \in \mathcal{J}_k \quad \text{with} \quad \ell \neq j(k). \quad (28b)$$

Then, setting the dual variables as

$$p_j^* = \left( \frac{1}{S_j} \sum_{k' \in \mathcal{K}_j} R_{k',j}^{\rho-1} \right)^{1/\rho}, \quad \text{and} \quad \lambda_k^* = 0, \quad (29)$$

we can easily verify that the KKT conditions (23) are satisfied with $\alpha_{k,j}^* = \frac{S_j R_{k,j}^{\rho-1}}{\sum_{k' \in \mathcal{K}_j} R_{k',j}^{\rho-1}}$ (from (26), yielding the user throughputs $r_k^* = \left( \frac{R_{k,j(k)}}{p_{j(k)}} \right)^{\rho}$ according to (22), which coincide with (27)). Since the KKT conditions are necessary and sufficient, we conclude that the valid partition $\{ \mathcal{K}_j : j \in \mathcal{J} \}$ combined with the local NUM policy (26) yields a globally optimal unique association configuration for the network-wide NUM problem (4). Also, we notice that the inequalities (28b) imply

$$\frac{S_j(k) R_{k,j}^{\rho}(k)}{\sum_{k' \in \mathcal{K}_j(k)} R_{k',j}^{\rho-1}(k)} \geq \frac{S_l R_{k,l}^{\rho}}{R_{k,l}^{\rho-1} + \sum_{k' \in \mathcal{K}_l} R_{k',l}^{\rho-1}}, \quad \forall \ k \in \mathcal{K}, \quad \text{and} \quad \ell \in \mathcal{J}_k \quad \text{with} \quad \ell \neq j(k). \quad (30)$$

This means that the throughput $r_k^*$ is larger than the “promised throughput” obtained if user $k$ changes unilaterally its association to some other BS $\ell \in \mathcal{J}_k$. Summarizing, we have proved:

**Lemma 2.** If a valid partition $\{ \mathcal{K}_j : j \in \mathcal{J} \}$ satisfies (28), then that valid partition is a Nash equilibrium of the decentralized association game which corresponds to a globally optimum unique association configuration for the network-wide NUM problem (4).

**Remark 3.** Notice that the condition (28a) determines a regime of “heavy-loaded” system, i.e., a system where the load $\sum_{k' \in \mathcal{K}_j} R_{k',j}^{\rho-1}$ of each BS is larger than its (scaled) spatial degrees of freedom $S_j R_{k,j}^{\rho-1}$. Otherwise, some users would have unit activity fractions, i.e., they would be served on all RBs. For a typical network serving a large number of users, this is the most relevant regime where all downlink data streams are used and non-trivial scheduling across the users must be performed.
An improvement path in the association game consists of a sequence of unique association configurations \( j \), each differing from the preceding one in a single component only, such that the change of strategy in that component increases the throughput of the corresponding user. If every improvement path is finite, then the game is said to have the finite improvement path property, which implies the existence of pure Nash equilibrium (see [16] and references therein).\(^7\)

Suppose that a pure Nash equilibrium exists. Hence, there exists a unique association (pure strategy) \( j \) such that (30) holds. Arguing in the reverse direction of the argument leading to Lemma 2, we observe that if the heavy-loaded condition (28a) is also verified, then the KKT conditions (23) of the global problem are “almost” satisfied. For example, in the case \( \gamma = 1 \) (PF), the Nash equilibrium conditions are given by

\[
\frac{S_{j(k)}R_{k,j(k)}}{|\mathcal{K}_{j(k)}|} \geq \frac{S_k R_{k,\ell}}{|\mathcal{K}_\ell| + 1}, \quad \forall \; k \in \mathcal{K}, \text{ and } \ell \in \mathcal{J}_k \text{ with } \ell \neq j(k).
\] (31)

When \( |\mathcal{K}_j| > S_j \gg 1 \) for all \( j \in \mathcal{J} \) (heavy-loaded system, typical in the case of massive MIMO networks), we have that the “+1” in the denominator of the promised rate terms is negligible with respect to the set size \( |\mathcal{K}_\ell| \), and therefore the Nash equilibrium condition and the KKT conditions (23) almost coincide. This means that, for heavy-loaded systems, a decentralized user-centric system operating at its Nash equilibrium is also very close to the global optimum of the network-wide NUM problem.

**B. User-centric decentralized on-line algorithms**

For a decentralized on-line implementation of the user-centric association scheme, several variants have been proposed.\(^8\) Here, and in our simulations, we restricted to a very simple scheme, that requires only local information at the users. In practice this can be easily obtained from the BSs through the “beacon-stuffing” approach [32], where each BS advertises the required information while broadcasting its beacon signal. In the proposed scheme, starting from a current association \( j \), each user \( k \) compares its current throughput \( r_k(j) \) given by (27) with the highest promised throughput

\[
\hat{r}_k = \max_{\ell \in \mathcal{J}_k \setminus j(k)} \min \left\{ \frac{S_k R_{k,\ell}^\rho}{\sum_{\ell' \in \mathcal{K}_k} R_{k',\ell'}^{\rho-1}}, R_{k,\ell} \right\}.
\] (32)

\(^7\)In general, the association game at hand is a finite game. Therefore, mixed strategy Nash equilibria are guaranteed to exist [31], but we cannot exclude a priori the case where pure Nash equilibria do not exist, when the conditions of Lemma 2 are not satisfied or the finite improvement path property cannot be proved.

\(^8\)For example, [16] discusses also the case where two subsets of BSs operate according different local NUM policy, a subset operates according to PF (equal air-time) and another subset operates according to HF (equal throughput).
If \( \hat{r}_k > r_k(j) \), user \( k \) changes its association to BS \( \hat{\ell} \) achieving the max in (32), with a certain probability \( \pi \in (0, 1) \). Otherwise, user \( k \) keeps its current association.

This algorithm evolves according to discrete-time finite-state Markov Chain, with state space \( J_1 \times \cdots \times J_K \). Since the Markov Chain has a finite state space and since every state has a self-transition of positive probability, the chain is aperiodic. Furthermore, it is immediate to see that the Nash equilibria of the game are absorbing states. Hence, if no pure Nash equilibrium exists the chain is ergodic and the scheme keeps evolving according to the above probabilistic best-response scheme. On the other hand, if all states \( j \) communicate with a Nash equilibrium state (i.e., there is a path of positive probability from \( j \) to a Nash equilibrium state), then the chain is decomposable, and the only persistent classes are the Nash equilibria, while all other states are transients. In this case, we have that the algorithm converges to a Nash equilibrium with probability 1. In particular, we have:

**Lemma 3.** If the association game has the finite improvement path property, for any switching probability \( \pi \in (0, 1) \) the randomized user-centric association algorithm converges to a Nash equilibrium.

*Proof:* For every user, if there exists a payoff improving BS, then the user switches to such a BS with a strictly positive probability \( \pi \). Otherwise, it maintains its current association with probability 1. Now, since the association game has the finite improvement path property, every unique association configuration communicates with a Nash equilibrium, i.e., starting from any unique association configuration, there exists a path of strictly positive probability to a Nash equilibrium state.

It is known from [16] that for the special cases of \( \gamma = 1 \) (PF) and \( \gamma \to \infty \) (HF), the association game has the finite improvement path property. Hence, in these cases, the user-centric decentralized association algorithm converges with probability 1. In our simulations, we noticed that convergence is always achieved also for \( 1 < \gamma < \infty \). We conjecture that pure Nash equilibria exist with high probability for random network topologies, and that convergence holds more in general. However, at this time, we are not able to prove or disprove such a conjecture.

**Remark 4.** In practice, networks have a (slowly) time-varying topology due to user motion across the coverage area. Hence, association must be continuously updated. We notice that the user-centric randomized scheme proposed here, with non-vanishing switch probability \( \pi \), is naturally suited to this purpose, allowing association switching as the user peak rates evolve in time due to user mobility. Also, it is practical to introduce some hysteresis in order to prevent too frequent association switches (which incur some protocol cost) and too wild fluctuations of the user peak rates. However, these practical
considerations go beyond the scope of this paper.

VI. SIMULATION RESULTS

In this section we present a comparative evaluation of the user-centric load-balancing scheme considered in this paper, and compare its performance against the optimal centralized solution of the NUM problem and a heuristic scheme based on maximum peak-rate association. In particular, there is no “standard” commonly accepted way to perform user-BS association in a network employing multiuser MIMO. Here, as a term of comparison we have chosen a naive maximum peak rate association scheme, i.e., user $k$ associates with BS $j(k) = \arg \max_{j \in \mathcal{J}} R_{k,j}$. In the massive MIMO regime, the user peak rates converge to the deterministic limit (1) of Section II-A which depends only on the individual user SINR (the term $\frac{P_j g_{jk}}{1 + \sum_{j' \neq j} P_{j'} g_{j'k}}$ in (1)) and on the BS specific spatial load factor (the term $\frac{M_j - S_j + 1}{S_j}$ in (1)), both of which can be assumed to be known. Hence, the Max peak rate association scheme can be easily implemented in the massive MIMO case. After the users associate with the BSs using the Max peak rate decision, the BSs locally implement the $\gamma$-fairness policy according to (26). We remark that the Max-RSSI scheme mentioned in Section I does not apply since, in general, the Signal to Interference plus Noise Ratio (SINR) achieved by any user with multiuser MIMO downlink spatial multiplexing depends on the channel matrix realization and on the set of simultaneously scheduled users.

Our simulations are based on a network topology formed by a 900m $\times$ 900m region with several small-cell BSs and a single macro BS whose locations are fixed throughout all of the simulation runs. As shown in Fig. 1a, the macro BS is located in the center (indicated by the $\circ$), and 20 small cell BSs (indicated by $\ast$’s) are uniformly distributed in the region. The number of users (indicated by $\ast$’s) and their locations change across different simulation runs, and are generated according to a non-homogeneous Poisson point process with higher density in a central region of size 300m $\times$ 300m, as shown in Fig. 1a. For the realization in Fig. 1a, Fig. 1b shows a visualization of the user-BS association for a randomly selected 20% percent of the users as provided by the centralized solution of (4) for $\gamma = 1$. An edge connecting BS $j$ and user $k$ indicates that $\alpha_{k,j}^* > 0$.

The macro BS has $M = 100$ antennas and serves user sets of size $S = 10$, with 46dBm transmission power. Each small-cell BS has $M = 40$ antennas and serves user sets of size $S = 4$, with transmission power of 35dBm. The pathloss from the macro BS to a user and from a small-cell BS to a user is given by $\frac{1}{1 + \left(\frac{d}{d_0}\right)^4}$ and by $\frac{1}{1 + \left(\frac{d}{d_0}\right)^4}$, respectively,\footnote{A greater pathloss exponent (4) is used for small-cell BSs, in order to take into account the fact that the macro-BS antennas are at higher elevation.} with $d$ representing the BS-user distance (assuming a torus

1. 1
wrap-around model to avoid boundary effects).

(a) Wireless network of small cells and a macro cell. (b) Illustration of the BS-user associations.

Fig. 1: Simulation setup

In Figs. 2a-2d and 3a-3d we compare the performance of the proposed centralized and distributed algorithms with the Max peak rate association scheme. For every realization of the layout similar to Fig. 1a, the network utility, the 5% percentile throughput, the geometric mean of user throughputs and the arithmetic mean of user throughputs are calculated and the CDFs of these quantities over 100 realizations are plotted for the case of $\gamma = 1$ (PF scheduling) in Figs. 2a, 2b, 2c and 2d respectively. We have run the centralized solution based on the dual subgradient algorithm of Section IV and also included results using the general purpose convex solver CVX, as a sanity check. Remarkably, the performance of the randomized distributed user-centric algorithm is almost indistinguishable from the performance of the (optimal) centralized solution, as we have argued to hold for highly-loaded systems in Section V. Furthermore, the figures reveal the fact that Max peak rate association results in inferior performance in terms of the 5% percentile throughput, the network utility and the geometric mean. Nevertheless, it achieves higher average throughput, since the PF fairness function imposes to serve all users in a proportionally fair way across the network, while Max peak rate does not.

Similar results can be observed for larger values of $\gamma$, where of course the larger the $\gamma$, the more fair the system behavior with respect to the user throughputs. For example, analogous results for $\gamma = 4$ are shown in Figs. 3a-3d. We remark that a higher 5% percentile throughput is achieved for larger $\gamma$, at the expense of lower geometric and arithmetic mean throughputs, reflecting that the system achieves better fairness across the users.
Fig. 2: Comparison of load balancing performance of various algorithms for $\gamma = 1$
Fig. 3: Comparison of load balancing performance of various algorithms for $\gamma = 4$
APPENDIX A

PHYSICAL REALIZABILITY OF FEASIBLE ASSOCIATION CONFIGURATIONS

**Theorem 1.** For any feasible association configuration \( \{\alpha_{k,j}\} \) (ref. Definition 4), there exists a physically realizable schedule, i.e., a sequence of integer scheduling configurations (ref. Definition 6) such that, by serving the users over a sequence of RBs according to such a sequence, the resulting long-term time averaged activity fractions \( \{\hat{\alpha}_{k,j}\} \) can be made arbitrarily close to the given feasible association configuration \( \{\alpha_{k,j}\} \).

**Proof:** We represent the network by a bipartite graph \( G = (J, K, E) \) where \( J \) is the set of BS nodes, \( K \) is the set of user nodes, and \( E = J \times K \) is the set of edges indicating possible association (for simplicity, here we let \( J_k = J \forall k \)). An integer scheduling configuration corresponds to a collection of edges \( F \subseteq E \), such that each BS \( j \in J \) is incident to at most \( S_j \) edges in \( F \), while each user \( k \) is incident to at most one edge in \( F \). When \( S_j = 1 \forall j \in J \), an integer scheduling configuration \( F \) corresponds to a matching in \( G \). For \( S_j > 1 \), we can think of an integer scheduling configuration \( F \) as a generalized matching. We now associate a point in \( \mathbb{R}^{|E|} \) to every integer scheduling configuration \( F \). For this purpose, given an integer scheduling configuration \( F \), let its incidence vector be \( \sigma \) where \( \sigma_{k,j} = 1 \) if \( (k, j) \in F \) and 0 otherwise. Let \( \Omega \) denote the set of incidence vectors where each incidence vector corresponds to an integer scheduling configuration. By sharing the RBs among such integer scheduling configurations, any feasible association configuration in the convex hull of \( \Omega \) can be achieved in the sense of long-term average. Let

\[
P' = \text{coh}(\Omega)
\]

(33)

denote the convex polytope obtained by taking the convex hull of the points in \( \Omega \). Also, let \( P \) denote the convex polytope corresponding to the set of linear constraints (4c)–(4e). The relation between the convex polytopes \( P \) and \( P' \) is not clear a priori. If one could show that \( P = P' \), then any feasible association configuration can be realized by first expressing the vector of user activity fractions as a convex combination of integer scheduling configurations in \( \Omega \), and then sharing the RBs among those configurations with scheduling dictated by the convex combination. Hence, proving \( P = P' \) implies the proof of Theorem 1. We shall prove this assertion by showing that both the relations \( P \subseteq P' \) and \( P' \subseteq P \) hold.

**Proposition 1.** \( P' \subseteq P \).
Proof: Consider any integer scheduling configuration \( \sigma \in \Omega \). It is easy to check that \( \sigma \) satisfies the constraints (4c)–(4e). Thus, \( \Omega \subseteq P \) holds and since \( P \) is a convex polytope, \( P' = \text{coh}(\Omega) \) is also a subset of \( P \).

Proposition 2. \( P \subseteq P' \).

We state a series of lemmas which provide a proof of Proposition 2. While we give proofs for certain lemmas, the other lemmas are well known and the reader is referred to the relevant literature in combinatorial optimization (see [33], [34] for example). The goal is to show that the set of extreme points (vertices) of \( P \) is included in the set of incidence vectors of integer scheduling configurations \( \Omega \). Once this is shown, we have our result since \( P = \text{coh}(\text{ext}(P)) \subseteq \text{coh}(\Omega) = P' \) where \( \text{ext}(P) \) is the set of the extreme points of \( P \). We re-write \( P \), i.e., the constraints (4c)–(4e) as:

\[
P = \{ \alpha : \begin{bmatrix} J & K \\ -I & 0 \end{bmatrix} \alpha \leq \begin{bmatrix} s \\ 1 \\ 0 \end{bmatrix} \} = \{ \alpha : A\alpha \leq b \},
\]

where \( A = \begin{bmatrix} J \\ K \\ -I \end{bmatrix} \) and \( b = \begin{bmatrix} s \\ 1 \\ 0 \end{bmatrix} \). Here, \( J \) is a matrix of dimensions \(|J| \times |E|\) with elements in the binary set \( \{0, 1\} \), where columns are indexed by edges in \( E \) and the rows are indexed by the BSs in \( J \). Each column has exactly one 1 corresponding to the BS on which the edge is incident. \( s \) is a \(|J| \times 1\) column vector with elements \( S_j \) \( \forall j \in J \). Similarly, \( K \) is a matrix of dimensions \(|K| \times |E|\) with elements in \( \{0, 1\} \), where columns are again indexed by the edges in \( E \) and the rows are indexed by the users in \( K \). Again, each column has exactly one 1, corresponding to the user on which the edge is incident. \( I \) is a \(|K| \times 1\) all-1 column vector. \( I \) is the \(|E| \times |E|\) identity matrix. Note that \( G = \begin{bmatrix} J \\ K \end{bmatrix} \) is the incidence matrix of the bipartite graph \( G \), i.e., \( G \) is a matrix of dimensions \((|J| + |K|) \times |E|\) and elements in \( \{0, 1\} \), where columns are indexed by edges in \( E \) and each column has exactly two 1’s, corresponding to the two vertices of the edge (one vertex in \( J \) and the other in \( K \)).

Definition 8. Extreme Point: A point \( \mathbf{v} \) in \( P \) is said to be an extreme point of \( P \) if it cannot be expressed as a convex combination of points in \( P \setminus \{\mathbf{v}\} \).

Let \( \mathbf{v} \) be an extreme point of the convex polytope \( P \) given by (34). Then, \( \mathbf{v} \) must satisfy the following lemmas:
Lemma 4. There are $|E|$ constraints in $A\alpha \leq b$ which are tight at $v$, i.e., $a_i^T v = b_i \forall i \in \{1, \ldots, |E|\}$ and in addition, $a_1, \ldots, a_{|E|}$ are linearly independent.

Proof: Consider $T = \{a_j : a_j^T v = b_j\}$. If $\dim(\text{span}(T)) < |E|$, then there exists $d \neq 0$ such that $d$ is orthogonal to $\text{span}(T)$, i.e., for all $a_j \in T$, $a_j^T d = 0$ and therefore $a_j^T (v \pm \epsilon d) = a_j^T v = b_j$. For all other constraints, $v$ satisfies strict inequality, i.e., $a_i^T v < b_i$, so there is some sufficiently small $\epsilon > 0$ such that $a_i^T (v + \epsilon d) \leq b_i$ and $a_i^T (v - \epsilon d) \leq b_i$. This means that $v + \epsilon d$ and $v - \epsilon d$ are in $P$ which in turn implies that $v = \frac{1}{2}(v + \epsilon d) + \frac{1}{2}(v - \epsilon d)$ is expressed as a convex combination of two other feasible points. This contradicts the fact that $v$ is an extreme point.

Lemma 5. $v$ is the unique solution to the $|E|$ constraints which are tight from Lemma 4.

Proof: The set of $|E|$ linear equations from Lemma 4 is a rank $|E|$ system of linear equations in $|E|$ dimensions. Thus, $v$ is the unique solution to the system.

It follows that every extreme point (or vertex) of $P$ is a unique solution to the linear system obtained from the tightness of $|E|$ constraints in the set of constraints $A\alpha \leq b$.

Definition 9. Totally Unimodular Matrix: A matrix $G$ is said to be totally unimodular if every square submatrix of $G$ has determinant $0$, $+1$ or $-1$.

Lemma 6. For all bipartite graphs $G$, the incidence matrix $G$ is totally unimodular.

Lemma 7. If $G$ is totally unimodular, then \[
\begin{bmatrix}
G \\
-I
\end{bmatrix}
\]
is totally unimodular.

See [33] for proofs of Lemmas 6 and 7. In particular, the incidence matrix $G = \begin{bmatrix} J \\ K \end{bmatrix}$ of the bipartite graph $G$ is totally unimodular from Lemma 6.

Let $v$ be a vertex of $P$. From Lemmas 4 and 5 there exists a rank $|E|$ square submatrix $A'$ of $A$ such that $A' v = b'$ and $v$ is the unique solution to the system.

Lemma 8. $v$ is an integer vector and is in the set of integer scheduling configurations $\Omega$.

Proof: $A'$ is a full rank square submatrix of \[
\begin{bmatrix}
G \\
-I
\end{bmatrix}
\]and since \[
\begin{bmatrix}
G \\
-I
\end{bmatrix}
\]
is totally unimodular from Lemma 7 we have that $\det A' = \pm 1$. Now by Cramer's rule, we have the $i$-th component $v_i$ of $v$ as:

\[
 v_i = \frac{\det(A'|b')}{\det(A')}
\] (35)
where $A'_i|b'$ is $A'$ with the $i$-th column replaced by $b'$. Note that $b$ has all integer elements, implying that $b'$ is an integer vector. Thus, with $b'$ being an integer vector and $\det(A') = \pm 1$, we conclude that $v_i$ is an integer. Now, given that $v$ is an integer vector and it satisfies (4c)–(4e), the only possible way for which this can happen is that $v$ is an integer scheduling configuration, i.e., $v \in \Omega$. This concludes the proof of Proposition 2.

**REFERENCES**

[1] “Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018.” [Online]. Available: http://goo.gl/1XYhqY

[2] H. Ishii, Y. Kishiyama, and H. Takahashi, “A novel architecture for LTE-B: C-plane/U-plane split and phantom cell concept,” in *IEEE Globecom Workshops*, 2012, pp. 624–630.

[3] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Trans. on Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, 2010.

[4] J. Hoydis, S. Ten Brink, and M. Debbah, “Massive MIMO: How many antennas do we need?” in *IEEE 49th Annual Allerton Conference on Communication, Control, and Computing*, 2011, pp. 545–550.

[5] H. Huh, G. Caire, H. Papadopoulous, and S. Ramprashad, “Achieving massive MIMO spectral efficiency with a not-so-large number of antennas,” *IEEE Trans. on Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, 2012.

[6] J. Hoydis, M. Kobayashi, and M. Debbah, “Green small-cell networks,” *IEEE Vehicular Technology Magazine.*, vol. 6, no. 1, pp. 37–43, 2011.

[7] V. Chandrasekhar, J. Andrews, and A. Gatherer, “Femtocell networks: a survey,” *IEEE Commun. Magazine.*, vol. 46, no. 9, pp. 59–67, 2008.

[8] “Quantenna 4 × 4 MIMO technology.” [Online]. Available: http://www.quantenna.com/4x4-mimo.html

[9] “Broadcom 6 × 6 MIMO press release.” [Online]. Available: http://goo.gl/zyLQ54

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

[11] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, “An overview of load balancing in HetNets: Old myths and open problems,” *arXiv preprint arXiv:1307.7779*, 2013.

[12] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, “User association for load balancing in heterogeneous cellular networks,” *IEEE Trans. on Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, 2013.

[13] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, “Optimizing client association in 60 GHz wireless access networks,” *arXiv preprint arXiv:1301.2723*, 2013.

[14] Y. Bejerano, S.-J. Han, and L. E. Li, “Fairness and load balancing in wireless LANs using association control,” in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*. ACM, 2004, pp. 315–329.

[15] L. Li, M. Pal, and Y. R. Yang, “Proportional fairness in multi-rate wireless LANs,” in *Proc. IEEE INFOCOM*, 2008, pp. 1004–1012.

[16] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, “RAT selection games in HetNets,” in *Proc. IEEE INFOCOM*, April 2013, pp. 998–1006.
[17] A. F. Molisch, *Wireless Communications*. Wiley, 2010, vol. 15.

[18] H. Huh, A. M. Tulino, and G. Caire, “Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling,” *IEEE Trans. on Information Theory*, vol. 58, no. 5, pp. 2911–2934, 2012.

[19] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, “Argos: Practical many-antenna base stations,” in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*. ACM, 2012, pp. 53–64.

[20] R. Rogalin, O. Bursalioglu, H. Papadopoulos, G. Caire, A. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, “Scalable synchronization and reciprocity calibration for distributed multiuser MIMO,” *IEEE Trans. on Wireless Comm.*, vol. 13, no. 4, pp. 1815–1831, 2014.

[21] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, “Multiple-antenna channel hardening and its implications for rate feedback and scheduling,” *IEEE Trans. on Inform. Theory*, vol. 50, no. 9, pp. 1893–1909, 2004.

[22] A. M. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communications*. Foundations and Trends in Communications and Information Theory, 2004, vol. 1, no. 1.

[23] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[24] S. Sesia, I. Toufik, and M. Baker, *LTE: the Long Term Evolution-From theory to practice*. Wiley, 2009.

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

[26] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.0 beta,” [http://cvxr.com/cvx](http://cvxr.com/cvx), Sep. 2013.

[27] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2009.

[28] S. Boyd and A. Mutapcic, “Subgradient methods,” Lecture notes of EE364b, Stanford University, Winter 2006-2007.

[29] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[30] N. R. Devanur, C. H. Papadimitriou, A. Saberi, and V. V. Vazirani, “Market equilibrium via a primal-dual algorithm for a convex program,” *Journal of the ACM*, vol. 55, no. 5, p. 22, 2008.

[31] J. F. Nash *et al.*, “Equilibrium points in n-person games,” *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[32] R. Chandra, J. Padhye, L. Ravindranath, and A. Wolman, “Beacon-stuffing: Wi-Fi without associations,” in *IEEE Workshop on Mobile Computing Systems and Applications, HotMobile*, 2007, pp. 53–57.

[33] J. Vondrak, “Polyhedral techniques in combinatorial optimization,” Lecture notes of CS369P, Stanford University, Fall 2010.

[34] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*. Springer, 2003, vol. 24.