Data Offloading in Load Coupled Networks: A Utility Maximization Framework

Chin Keong Ho, Di Yuan and Sumei Sun

Chin Keong Ho, Di Yuan and Sumei Sun, Data Offloading in Load Coupled Networks: A Utility Maximization Framework, 2014, IEEE Transactions on Wireless Communications, (13), 4, 1921-1931.

DOI: http://dx.doi.org/10.1109/TWC.2014.021214.130809

Data Offloading in Load Coupled Networks: A Utility Maximization Framework

Chin Keong Ho, *Member, IEEE*, Di Yuan, *Member, IEEE*, and Sumei Sun, *Senior Member, IEEE*

***Abstract*—We provide a general framework for the problem of data offloading in a heterogeneous wireless network, where some demand of cellular users is served by a *complementary network*. The complementary network is either a small-cell network that shares the same resources as the cellular network, or a WiFi network that uses orthogonal resources. For a given demand served in a cellular network, the load, or the level of resource usage, of each cell depends in a non-linear manner on the load of other cells due to the mutual coupling of interference seen by one another. With load coupling, we optimize the demand to be served in the cellular or the complementary networks, so as to maximize a utility function. We consider three representative utility functions that balance, to varying degrees, the revenue from serving the users vs the user fairness. We establish conditions for which the optimization problem has a feasible solution and is convex, and hence tractable to numerical computations. Finally, we propose a strategy with theoretical justification to constrain the load to some maximum value, as required for practical implementation. Numerical studies are conducted for both under-loaded and over-loaded networks.**

***Index Terms*—Data offloading, load coupling, small-cell network, WiFi network, feasibility, convexity.**

I. Introduction

FUELED by mobile multimedia applications, the demand for mobile data is rising rapidly. Data traffic is also projected to grow at a compound annual growth rate of 78% from 2011 to 2016 [1]. In practice, cellular networks and the conventional infrastructure cannot grow as fast to match the increase in demand. One promising solution currently considered by cellular operators is to employ data offloading, also known as mobile cellular traffic offloading [2], [3]. In data offloading, the data of cellular users is intentionally delivered by complementary networks, namely small cells such as Picocells and Femtocells, or WiFi networks. This reduces the data demand on the regular cellular networks and hence eases traffic congestion.

In a cellular network, frequency reuse is employed, and thus base stations using the same frequency band interfere with one another. We refer to the average level of resource usage in the time-frequency domain of a cell as its load. To optimize the overall system performance, load balancing has to be performed across various networks in the context of data offloading [4]. Due to the mutual coupling of the interference and the requirement to serve a specific demand for each cell, the load of a cell depends on the load of other cells. This leads to a non-linear coupling relation of the cells' loads, making analytical characterization of the load challenging. This motivates the use of new approaches and different theoretical tools to analyze and optimize the system performance.

Recently, an analytical signal-to-interference-and-noise-ratio (SINR) model that takes into account the load of each cell is employed [5], [6], resulting in a non-linear load coupling equation for which theoretical analysis is obtained in [7]. This load coupling equation has also been shown to give a good approximation for more complicated load models in cellular systems that capture the dynamic nature of arrivals and service periods of data flows in the network, especially at high data arrival rates [8]. For example, the load obtained by the load coupling equation is within 10% of the exact load obtained, if the normalized arrival rate of the data is more than 60%.

In this paper, we consider two separate scenarios in which the cellular network offloads to a complementary network. The complementary network is either a small cell or a WiFi network. The small-cell network shares the network resources with the cellular network, whereas the WiFi network uses orthogonal network resources to that of the cellular network. The performance of serving users in a particular network is measured by three representative types of utility functions, all of which are related to the network operator's revenue, but differ in the degree of accounting for user fairness. To model the inter-dependency of the load, we employ the load coupling equation in [5]–[8].

In this paper, we extend the theoretical insights in [9], as well as present new algorithmic solutions and results for utility maximization with data offloading. Our contributions are as follows. Based on a unified framework for the problem of data offloading, we obtain fundamental properties on the computation, feasibility, and monotonicity of the load-coupling system. For a given (small cell or WiFi) complementary network, we formulate a utility-maximization problem in which the users' demand can be served in either the (regular) cellular network or the complementary network, or concurrently in both networks. We establish conditions for which the optimization problem has a feasible solution and is convex, and hence tractable to numerical computations. We also propose a strategy to constrain the load to some maximum value, as required for practical implementation, and provide theoretical justification for the proposed algorithm. Numerical results are obtained for both under-loaded and over-loaded networks, and can serve as a reference for the design of data offloading systems in practice. The main tool we employ for analysis is based on the Perron-Frobenius theorem and other related

This paper is presented in part at the IEEE International Conference on Communications, June 2013.

C. K. Ho and S. Sun are with the Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632 (e-mail: {hock, sunsm}@i2r.a-star.edu.sg).

D. Yuan is with the Department of Science and Technology, Linköping University, Sweden. (e-mail: di.yuan@liu.se)

The work of D. Yuan has been supported by the Swedish ELLIIT Excellence Center, CENIIT of Linköping University, Sweden, and the European FP7 Marie Curie IAPP scheme with contract number 324515.

results [10].

Section II gives the system model of the load-coupled network. Section III presents the fundamental properties of the load-coupled system. Section IV formulates the data offloading problem. Convexity analysis and an algorithm to constrain the maximum load are also given. Numerical results are given in Section V. Section VI concludes the paper.

Notations: We denote a (tall) vector by a bold lower case letter, say $\boldsymbol{a}$. We denote a matrix by a bold capital letter, say $\boldsymbol{A}$, and denote its (i,j)th element by its lower case a_{ij}. We denote a *positive* matrix as $\boldsymbol{A} > 0$ if $a_{ij} > 0$ for all i, j. Similarly, we denote a *non-negative* matrix as $\boldsymbol{A} \geq 0$ if $a_{ij} \geq 0$ for all i, j. Similar definitions apply to vectors.

II. System Model

We consider a cellular network consisting of n base stations that can interfere with each other. We focus on the downlink communication scenarios where base station $i \in \mathcal{N} \triangleq \{1, \cdots, n\}$ transmits with power $p_i \geq 0$. We refer to cell i interchangeably with base station i. For notational convenience, we collect all power $\{p_i\}$ as vector $\boldsymbol{p} > 0$.

Each base station i serves one unique group of users in set $\mathcal{J}_i$, where $|\mathcal{J}_i| \geq 1$. User $j \in \mathcal{J}_i$ is served in cell i up to a maximum rate of D_{ij} nat. Thus, the data can be interpreted as best-effort or elastic data to be served as much as possible subject to network conditions. We also allow the users to be served in a *complementary cell*, to be introduced next.

A. Data Offloading to Complementary Network

We shall consider data offloading, where the demand of every user can also be served in a complementary network. We assume a total of n' complementary cells in the complementary network, denoted by the set $\mathcal{N}' = \{1, \cdots, n'\}$. Each complementary cell i transmits with power $p'_i \geq 0$.

Specifically, we map every regular-cell user $j \in \mathcal{J}_i$ in the regular cell $i \in \mathcal{N}$ uniquely to a (virtual) complementary-cell user $b \in \mathcal{J}'_a$ in the complementary cell $a \in \mathcal{N}'$, via the mapping $(a, b) = \pi(i, j)$; note that both refers to the same physical user. We take the demands d_{ij} and $d'_{\pi(i,j)}$ to be served in the regular and complementary cells, respectively, as variables to be optimized, subject to the demand constraint

$$d_{ij} + d'_{\pi(i,j)} \leq D_{ij}, \quad i \in \mathcal{N}, j \in \mathcal{J}_i. \tag{1}$$

The demand constraint ensures that the total demand served to each user is not more than the demand D_{ij} requested. This is because any demand served beyond the requested amount may not benefit the users, yet consumes additional network resources at an increased cost for the cellular operator. For notational convenience, we collect all demands $\{d_{ij}\}$ and $\{d'_{\pi(i,j)}\}$ as vectors $\boldsymbol{d} \geq 0$ and $\boldsymbol{d}' \geq 0$, respectively.

We assume there is at least one user j in cell i with $d_{ij} > 0$, otherwise $p_i = 0$ and so base station i can be omitted; we make the same assumption for the complementary cells. Thus without loss of generality, we have $p_i, p'_i > 0$.

We consider two types of complementary network, consisting of either only small cells or WiFi cells. For the case of small-cell offloading, both the regular cellular network and small-cell network use the same frequency band, hence the networks interfere with each other. For the case of WiFi offloading, the frequency band used in the WiFi network is orthogonal to that of the cellular network, hence there is no mutual interference. Our model can be easily generalized to the hybrid case consisting of a mixture of small cells and WiFi cells, with more cumbersome notations. For ease of exposure, we do not consider this hybrid case.

B. Load Coupling Model

We first consider the load coupling model for the cellular network without any complementary network. The extension to the case with a complementary network is given in Section II-C.

Let $\boldsymbol{x} = [x_1, \cdots, x_n]$ be the load of the cellular network, where $0 \leq \boldsymbol{x} \leq 1$. The load x_i measures the fractional usage of resource in cell i. In LTE systems, the load can be interpreted as the expected fraction of the time-frequency resources that are scheduled to deliver data. We model the SINR of user j in cell i as [5]–[8]

$$\mathsf{SINR}_{ij}(\boldsymbol{x}) = \frac{p_i g_{ij}}{\sum_{k \in \mathcal{N} \setminus \{i\}} p_k g_{kj} x_k + \sigma^2} \tag{2}$$

where σ^2 represents the noise power and g_{ij} is the channel gain (or channel power) from base station i to user j; note that $g_{kj}, k \neq i$, here represents the channel gain from interfering base station k. The SINR model (2) gives good approximation of more complicated cellular models [8]. Intuitively, x_k can be interpreted as the probability of receiving interference from cell k on all the sub-carriers of the resource unit. Thus, the combined term $(p_k g_{kj} x_k)$ is interpreted as the expected interference with expectation taken over time and frequency for all transmissions.

Since Gaussian-signalling is the worst-case noise distribution for mutual information [11], an achievable rate is given by $r_{ij} = B \log(1 + \mathsf{SINR}_{ij})$ nat/s per resource unit, where B is the bandwidth for one resource unit and $\log$ is the natural logarithm. To deliver a demand of d_{ij} nat for user j, the ith base station thus uses $x_{ij} \triangleq d_{ij}/r_{ij}$ resource units. We assume that at total M (time and frequency) resource units are available. Summing the resource units over all users in cell i, we get the load for the cell as

$$x_i = \sum_{j \in \mathcal{J}_i} x_{ij}/M \tag{3}$$

$$= \frac{1}{MB} \sum_{j \in \mathcal{J}_i} \frac{d_{ij}}{\log\left(1 + \mathsf{SINR}_{ij}(\boldsymbol{x})\right)} \triangleq f_i(\boldsymbol{x}) \tag{4}$$

for $i \in \mathcal{N}$. For notational simplicity, we normalize d_{ij} and r_{ij} by the total amount of resource units MB. Hence, without loss of generality we let $MB = 1$ in (4).

Let $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x})]^T$. In vector form, we have

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{d}, \boldsymbol{p}) \tag{5}$$

where we have made the dependence of the load on the demand $\boldsymbol{d}$ and power $\boldsymbol{p}$ explicit. We call (5) the *non-linear load coupling equation* (NLCE), as the load $\boldsymbol{x}$ appears in both sides of the equation and cannot be readily solved in closed-form. To emphasize that a load is a solution of the NLCE, we denote the load as $\boldsymbol{x}^\star$ when necessary. We say the load

$\boldsymbol{x}^\star$ to be *feasible* if $\boldsymbol{x}^\star$ satisfies the NLCE *and* $\boldsymbol{x}^\star \geq 0$. An algorithm that ensures that the load is also less than one (by reducing the demand) shall be considered in Section IV-E.

C. Load Coupling with Complementary Network

For the cellular network with a small-cell network, the two networks operate in the same frequency band and can be treated as one integrated network. Specifically, the set of n base stations in the regular network is combined with the set of n' base stations in the small-cell network to form a larger set of base stations of size $n+n'$. All base stations can interfere with one another.

For the cellular network with a WiFi network, the two networks operate in different frequency bands. We assume the WiFi network also submits to the load-coupling system relation. That is, the NLCE holds for the cellular network as before, and also holds separately for the WiFi network by replacing $\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{p}$ with the corresponding WiFi quantities denoted by $\boldsymbol{x}', \boldsymbol{d}', \boldsymbol{p}'$.

We note that regardless of whether the complementary network is a small cell or WiFi network, the allocation of $\{d_{ij}, d'_{\pi(ij)}\}$ is coupled due to the constraint $d_{ij} + d'_{\pi(ij)} \leq D_{ij}$.

III. Feasible Load: Fundamental Properties

We explore fundamental properties related to NLCE, namely, computation, existence and monotonicity of the load solution. For clarity, we consider the regular cellular network without any complementary network; the results extend straightforwardly to the case with complementary network via the discussion in Section II-C.

A. Computation

Consider the following iterative algorithm. Starting from an arbitrary initial load $\boldsymbol{x}^0 > 0$, define the kth iteration solution as

$$\boldsymbol{x}^k = \boldsymbol{f}(\boldsymbol{x}^{k-1}; \boldsymbol{d}, \boldsymbol{p}) \tag{6}$$

for $k = 1, 2, \cdots, K$, where K is the total number of iterations. Lemma 1 ensures that $\boldsymbol{x}^K$ converges to the feasible load $\boldsymbol{x}^\star$ in the NLCE for large K. The proof relies on the property of the standard interference function as defined in [12].

Lemma 1: Suppose a feasible load $\boldsymbol{x}^\star$ exists for the NLCE (5). Then $\boldsymbol{x}^K$ converges to the unique fixed point solution $\boldsymbol{x}^\star$ as $K \to \infty$.

Proof: We sketch the proof given in [8]. After establishing that $\boldsymbol{f}(\cdot)$ is a standard interference function, Theorem 2 in [12] is applied to obtain the desired result. ∎

Remark 1 (Asynchronous iteration): The iterative algorithm in (6) is said to be *synchronous* [12] because all elements in vector $\boldsymbol{x}^k$ are obtained simultaneously. We may also consider the *asynchronous* version, in which a set of one or more elements are updated multiple times followed sequentially by other sets until all cells are updated at least once. By using Theorem 4 in [12], we also obtain the convergence property in Lemma 1 with asynchronous iterations.

The observation in Remark 1 is useful for implementation in practice, because the base stations can adapt their load in a distributed manner, and yet a feasible load can be obtained after sufficient number of iterations. Moreover, the asynchronous iteration will be used from an analytical viewpoint later, in the proof of Theorem 2.

B. Existence

Before we compute the load as in Lemma 1, we need to check if a feasible load exists. Lemma 2 next states that feasibility can be checked by a simpler problem via a *linear* counterpart to the NLCE.

Lemma 2: Given $\boldsymbol{d}$ and $\boldsymbol{p}$, a feasible load $\boldsymbol{x}^\star \geq 0$ exists for the NLCE (5) if and only if a solution $\boldsymbol{x} \geq 0$ exists in

$$\boldsymbol{x} = \boldsymbol{H}(\boldsymbol{d}, \boldsymbol{p}) \cdot \boldsymbol{x} + \boldsymbol{c}(\boldsymbol{d}, \boldsymbol{p}). \tag{7}$$

Here, $\boldsymbol{c}(\boldsymbol{d}, \boldsymbol{p}) \triangleq \boldsymbol{f}(\boldsymbol{0}_n; \boldsymbol{d}, \boldsymbol{p})$, where $\boldsymbol{0}_n$ is the length-n all-zero vector, and $\boldsymbol{H}(\boldsymbol{d}, \boldsymbol{p}) \geq 0$ is the real matrix with (i,k)th element

$$h_{ik} = \begin{cases} 0, & \text{if } i = k; \\ (p_k/p_i) \sum_{j \in \mathcal{J}_i} g_{kj} d_{ij} / g_{ij}, & \text{if } i \neq k \end{cases} \tag{8}$$

for $1 \leq i \leq n$ and $1 \leq k \leq n$. Note that $\boldsymbol{c}(\boldsymbol{d}, \boldsymbol{p}) > 0$ because at least one d_{ij} in cell i is positive.

Proof: From Theorem 8 and Theorem 11 in [7]. ∎

Next, we treat $\boldsymbol{d}$ and $\boldsymbol{p}$ as variables to be optimized, so as to study how they affect the feasibility of the load. Our main result is stated in Theorem 1 below, which gives the necessary and sufficient condition for a feasible $\boldsymbol{x}^\star$ to exist.

We make some preparation before stating the theorem. Let $\boldsymbol{\Lambda}(\boldsymbol{d}) \geq 0$ be the n-by-n real matrix with the (i,k)th element

$$\lambda_{ik} = \begin{cases} 0, & \text{if } i = k; \\ \sum_{j \in \mathcal{J}_i} g_{kj} d_{ij} / g_{ij}, & \text{if } i \neq k \end{cases} \tag{9}$$

for $1 \leq i \leq n$ and $1 \leq k \leq n$. We can therefore express the matrix $\boldsymbol{H}(\boldsymbol{d}, \boldsymbol{p})$ in (7) as

$$\boldsymbol{H}(\boldsymbol{d}, \boldsymbol{p}) = \mathrm{diag}(\boldsymbol{p}) \cdot \boldsymbol{\Lambda}(\boldsymbol{d}) \cdot \mathrm{diag}(\boldsymbol{p})^{-1} \tag{10}$$

where $\mathrm{diag}(\boldsymbol{p})$ denotes the diagonal matrix with diagonal elements $\boldsymbol{p}$. The effects of $\boldsymbol{p}$ and $\boldsymbol{d}$ are thus decoupled into three matrices, and so (7) becomes

$$\widetilde{\boldsymbol{x}} = \boldsymbol{\Lambda}(\boldsymbol{d}) \widetilde{\boldsymbol{x}} + \widetilde{\boldsymbol{c}}(\boldsymbol{p}, \boldsymbol{d}) \tag{11}$$

where $\widetilde{\boldsymbol{x}} \triangleq \mathrm{diag}(\boldsymbol{p})^{-1} \boldsymbol{x}$ and $\widetilde{\boldsymbol{c}}(\boldsymbol{d}, \boldsymbol{p}) \triangleq \mathrm{diag}(\boldsymbol{p})^{-1} \boldsymbol{c}(\boldsymbol{d}, \boldsymbol{p})$.

Theorem 1: Given $\boldsymbol{p} > 0$ and $\boldsymbol{d} \geq 0$, a feasible load $\boldsymbol{x}^\star \geq 0$ for the NLCE (5) exists if and only if

$$r(\boldsymbol{\Lambda}(\boldsymbol{d})) < 1 \tag{12}$$

where $r(\boldsymbol{\Lambda})$ is the *spectral radius* of matrix $\boldsymbol{\Lambda}$, defined as the absolute value of the largest eigenvalue of $\boldsymbol{\Lambda}$.

Proof: By Lemma 2, it is sufficient to consider the linear counterpart (7), or equivalently (11). Since $\boldsymbol{p} > 0$, every base station i serves some positive demand and so $\sum_{j \in \mathcal{J}_i} d_{ij} > 0$. Thus, $\boldsymbol{\Lambda}(\boldsymbol{d}) \geq 0$ and $\boldsymbol{c}(\boldsymbol{d}, \boldsymbol{p}) > 0$. Hence, applying the Perron-Frobenius theorem in [10, Theorem A.51] to (11), we conclude that (12) is necessary and sufficient for a feasible $\widetilde{\boldsymbol{x}}$ to exist in (11). Theorem 1 follows as $\boldsymbol{p} > 0$. ∎

From (12), the existence of a feasible load depends only on the demand vector $\boldsymbol{d}$, but not on the power $\boldsymbol{p}$. This suggests the

importance of data offloading by varying the demand, which is made explicit in Corollary 1.

Corollary 1: Suppose a feasible load does not exist for a given demand $\boldsymbol{d} \geq 0$ and power $\boldsymbol{p} > 0$. Then no feasible load can exist by varying only $\boldsymbol{p}$. However, a feasible load always exist by varying $\boldsymbol{d}$.

Proof: The spectral radius $r(\boldsymbol{\Lambda}(\boldsymbol{d}))$ depends only on the demand $\boldsymbol{d}$. Hence, changing the power $\boldsymbol{p}$ does not affect the existence of the feasible load. But scaling the demand vector uniformly by a positive factor allows the spectral radius to be scaled also by the same factor. Hence the spectral radius can always be made smaller than one by reducing the demand such that a feasible load exists. ∎

Motivated by Corollary 1, subsequently we shall focus on the scenario where only the demand is varied, while the power is always taken to be fixed and positive.

C. Monotonicity of Load as a Function of Demand

With power fixed, Theorem 2 shows that the load vector that satisfies the NLCE is a monotonic function of the demand vector.

Theorem 2: Consider the NLCE (5) with power $\boldsymbol{p}$ fixed. Given the demand vectors $\boldsymbol{d}'$ and $\boldsymbol{d}$ with $\boldsymbol{d}' \geq \boldsymbol{d}$ and $\boldsymbol{d}' \neq \boldsymbol{d}$, the corresponding NPCE load $\boldsymbol{x}'^{\star}$ and $\boldsymbol{x}^{\star}$ satisfy $\boldsymbol{x}'^{\star} > \boldsymbol{x}^{\star}$.

Proof: We sketch the proof; the details are given in Appendix A. First, consider the case that only one element of $\boldsymbol{d}'$ is strictly greater than $\boldsymbol{d}$, with all other demand elements unchanged. Then we employ an asynchronous iteration in Remark 1 with initial load $\boldsymbol{x}^{\star}$. Upon convergence, we obtain $\boldsymbol{x}'^{\star}$ which can be shown to satisfy $\boldsymbol{x}'^{\star} > \boldsymbol{x}^{\star}$. Finally we consider the case where more than one element of $\boldsymbol{d}'$ are strictly greater than $\boldsymbol{d}$. In the proof, we apply the above argument successively to each element of the demand vector where the strict inequality holds. ∎

Theorem 2 also justifies our approach of focusing on a feasible load vector such that $\boldsymbol{x}^{\star} \geq 0$. Once feasibility is established, we can reduce the demand further to ensure that $0 \leq \boldsymbol{x}^{\star} \leq 1$.

IV. Demand Offloading

We model the benefit of serving the demand in a network with offloading via three representative utility functions in Section IV-A. Next, we pose the optimization problem of maximizing the sum utility in Section IV-B, where the complementary network is either a WiFi or small-cell network. Then we investigate the convexity of the solution space, which affects the difficulty of numerical computations of the optimal solution, for $n = 2$ base stations in Section IV-C and for arbitrary n in Section IV-D. Finally, we propose an algorithm to limit the maximum optimal load to one in Section IV-E.

As before, we shall consider feasible load such that $\boldsymbol{x}^{\star} \geq 0$ in this section. In Section IV-E, we shall impose the additional constraint that the feasible load is less than one, i.e., $0 \leq \boldsymbol{x}^{\star} \leq 1$.

A. Utility for Maximization

Our objective is to maximize the sum utility

$$U^{\mathrm{sum}} \triangleq \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{J}_i} k_{ij} U(d_{ij}) + k'_{\pi(i,j)} U(d_{\pi(i,j)}) \quad (13)$$

where $U(d)$ is the utility function for satisfying demand d. The positive weights k_{ij} and $k'_{\pi(i,j)}$ take into account the combined priority of the user and the networks. The utility function can be used to quantify the value of serving the demand d to the cellular operator or user in terms of, for instance, the revenue collected from the access service, and the fairness of serving the demand of multiple users within each cell type. We note that the importance of serving in either cell type can be quantified via the weights k_{ij} and $k'_{\pi(i,j)}$.

To give insights, $U(d)$ is chosen to be the following representative functions, namely the linear (LIN), logarithmic (LOG), and double-logarithmic (DLOG) utility functions:

$$\mathsf{LIN}: U(d) = d, \quad (14a)$$

$$\mathsf{LOG}: U(d) = \log(d), \quad (14b)$$

$$\mathsf{DLOG}: U(d) = \log(\log(1+d)). \quad (14c)$$

The utility functions are monotonically increasing and hence one-to-one functions. The LIN utility models the scenario where serving an additional demand unit results in an additional unit of utility. For LOG utility, serving an additional demand unit of a user with a low demand results in more utility. Intuitively, this results in a fairer demand distribution among users but could result in a smaller revenue to the operator as less total demand is served. Thus, the LOG utility trades revenue maximization with user fairness. The DLOG utility further emphasizes fairness, because it favours low-demand users even more. We note that the last two utility functions would not assign zero demand to any user, because the sum utility is then negative infinity. The generalization to a broader class of functions is considered in Remark 3 later.

For exposure, we make the *same-demand assumption* that every user j in the same regular cell i is served the same demand $d_{ij} = \widetilde{d}_i$. Corresponding to the regular-cell user j in cell i, we denote the complementary-cell user as $a(i,j)$ in complementary cell $b(i,j)$, i.e., $(a(i,j), b(i,j)) = \pi(i,j)$. For the complementary network, we also make the same-demand assumption, i.e., $d'_{\pi(i,j)} = \widetilde{d}'_{a(i,j)}$ for all i, j. In effect, we focus on varying the cell-level demand vectors $\widetilde{\boldsymbol{d}} \triangleq [\widetilde{d}_1, \cdots, \widetilde{d}_n]^T$ and $\widetilde{\boldsymbol{d}}' \triangleq [\widetilde{d}'_1, \cdots, \widetilde{d}'_{n'}]^T$. From the demand constraint (1), we get

$$d_{ij} + d_{\pi(i,j)} = \widetilde{d}_i + \widetilde{d}_{a(i,j)} \leq D_{ij}, \forall j \in \mathcal{J}_i, i \in \mathcal{N}. \quad (15)$$

Since all cells are active with power vector $\boldsymbol{p} > 0$, we also have $\widetilde{\boldsymbol{d}} > 0$ and $\widetilde{\boldsymbol{d}}' > 0$.

Remark 2 (Relaxing same-demand assumption): For the case of LOG utility, the same-demand assumption can be slightly relaxed. Instead we assume more generally that each user $j \in \mathcal{J}_i$ in cell i is allocated a demand of $d_{ij} = \alpha_{ij}\widetilde{d}_i$, where $\sum_{j \in \mathcal{J}_i} \alpha_{ij} = 1$. Here, α_{ij} is a fraction of the total demand $\widetilde{d}_i$ served in cell i. Thus, the user's achieved utility is $U(\alpha_{ij}\widetilde{d}_i) = \log(\widetilde{d}_i) + \log(\alpha_{ij})$. With α_{ij}'s fixed, it suffices to consider the first term $\log(\widetilde{d}_i)$ for the sum utility U^{sum}. Hence the optimization problem is similar to the case under the same-demand assumption. In general, however, the same-demand assumption is required for the subsequent convexity results to hold.

B. Optimization Problem

1) WiFi as Complementary Network: We first formulate the optimization problem with WiFi as the complementary network. Mathematically, our *data offloading problem* is

$$(P0)\quad \max_{\widetilde{\boldsymbol{d}},\widetilde{\boldsymbol{d}}'} \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{J}_i} k_{ij}U(\widetilde{d}_i) + k'_{\pi(i,j)}U(\widetilde{d}'_{a(i,j)}) \quad (16a)$$

$$\text{s.t.}\quad \widetilde{\boldsymbol{d}} \in \mathcal{F} \triangleq \{\widetilde{\boldsymbol{d}} > 0 : r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}})) < 1\} \quad (16b)$$

$$\widetilde{\boldsymbol{d}}' \in \mathcal{F}' \triangleq \{\widetilde{\boldsymbol{d}}' > 0 : r(\boldsymbol{\Lambda}'(\widetilde{\boldsymbol{d}}')) < 1\} \quad (16c)$$

$$\widetilde{d}_i + \widetilde{d}'_{a(i,j)} \le D_{ij},\ \forall j \in \mathcal{J}_i, i \in \mathcal{N} \quad (16d)$$

where $\boldsymbol{\Lambda}, \boldsymbol{\Lambda}'$ correspond to (9) for the regular cellular network and the WiFi network, respectively. The constraints (16b), (16c) follow from Theorem 1 and the discussion for WiFi network in Section II-C; we call $\mathcal{F}$ and $\mathcal{F}'$ the *feasibility sets*. The last constraint is due to (15). We note that a solution always exists because we can always reduce the demand to arbitrarily close to zero, so as to satisfy constraints (16b) and (16c) (see the proof of Corollary 1) and also to satisfy constraint (16d).

For convenience, we transform $\widetilde{d}_i$ to $y_i = U(\widetilde{d}_i)$ for $i \in \mathcal{N}$ and let $\widetilde{\boldsymbol{d}} = [U^{-1}(y_1), \cdots, U^{-1}(y_n)]^T \triangleq \boldsymbol{g}(\boldsymbol{y})$. The inverse $U^{-1}(\cdot)$ always exists because $U(\cdot)$ is a monotonic function. Similarly for the WiFi cells, let $y'_i = U(\widetilde{d}'_i)$ for $i \in \mathcal{N}$ and $\widetilde{\boldsymbol{d}}' = [U^{-1}(y'_1), \cdots, U^{-1}(y'_n)]^T \triangleq \boldsymbol{g}(\boldsymbol{y}')$. Let $k_i \triangleq \sum_{j\in\mathcal{J}_i} k_{ij}$. We make similar definitions for y'_i and k'_i corresponding to the complementary cells. Our *transformed data offloading* problem is then

$$(P1)\quad \max_{\boldsymbol{y},\boldsymbol{y}'} \sum_{i\in\mathcal{N}} k_i y_i + k'_i y'_i \quad (17a)$$

$$\text{s.t.}\quad \boldsymbol{y} \in \widetilde{\mathcal{F}} \triangleq \{\boldsymbol{y} \in \mathcal{Y}^n : r(\boldsymbol{\Lambda}(\boldsymbol{g}(\boldsymbol{y}))) < 1\} \quad (17b)$$

$$\boldsymbol{y}' \in \widetilde{\mathcal{F}}' \triangleq \{\boldsymbol{y}' \in \mathcal{Y}^{n'} : r(\boldsymbol{\Lambda}'(\boldsymbol{g}(\boldsymbol{y}'))) < 1\} \quad (17c)$$

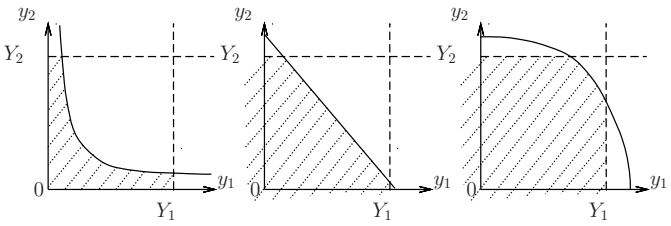
$$U^{-1}(y_i) + U^{-1}(y'_{a(i,j)}) \le D_{ij},\ \forall j \in \mathcal{J}_i, i \in \mathcal{N} \quad (17d)$$

where $\mathcal{Y}^n$ of dimension n is defined as the set of positive vectors for LIN utility, and as the set of real vectors for LOG and DLOG utilities. Here $\widetilde{\mathcal{F}}$ (and similarly $\widetilde{\mathcal{F}}'$) is the transformed feasibility set with complement set denoted as $\widetilde{\mathcal{F}}^c = \mathcal{Y}^n \setminus \widetilde{\mathcal{F}}$. For LIN utility, Problem $P1$ is the same as Problem $P0$, and thus $\widetilde{\mathcal{F}} = \mathcal{F}$.

We note that the objective function is always linear in $\boldsymbol{y}$ and $\boldsymbol{y}'$. For any of the three utility functions, it can be checked that the set of $\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{y}}'$ subject only to (17d) is convex. Now if $\widetilde{\mathcal{F}}$ (and similarly $\widetilde{\mathcal{F}}'$) is a convex set, $P1$ is a convex optimization problem for which numerically efficient solvers exist [13]. In summary, it is sufficient to obtain conditions for which $\widetilde{\mathcal{F}}$ is convex, in order to ascertain if problem $P1$ is convex.

To account for individual demand constraints imposed by each network, we may also impose, in Problem $P0$, additional constraints on $\widetilde{d}_i$ and $\widetilde{d}'_{a(i,j)}$ such that the variables do not exceed some fixed constants. It can be easily checked that such constraints do not affect the convexity of the solution space in Problem $P1$.

2) Small cell as Complementary Network: The optimization problem is similar to problem $P0$, except that the feasibility sets in (16b) and (16c) are merged into a single feasibility set subject to $r(\boldsymbol{\Lambda}''(\widetilde{\boldsymbol{d}}, \widetilde{\boldsymbol{d}}')) < 1$ where $\boldsymbol{\Lambda}''$ includes the base stations of both the regular cells and the small cells, see Section II-C. By similar arguments as before, for the transformed data offloading problem to be convex, it suffices to check if $\widetilde{\mathcal{F}}$ that corresponds to $\boldsymbol{\Lambda}''(\widetilde{\boldsymbol{d}}, \widetilde{\boldsymbol{d}}')$ is convex.



(a) LIN utility. (b) LOG utility. (c) DLOG utility.

Fig. 1. Transformed feasibility set (shaded) for different utility objective functions. After transformation, the objective function is always linear.

C. Two Base Stations

To gain some understanding for the convexity of $\widetilde{\mathcal{F}}$, let us study the case of $n = 2$ base stations. It can be verified that if we write $\boldsymbol{\Lambda}(\boldsymbol{d}) = \begin{bmatrix} 0 & \beta \\ \beta' & 0 \end{bmatrix}$, then the unit eigenvectors and corresponding eigenvalues of $\boldsymbol{\Lambda}(\boldsymbol{d})$ are $\left\{\xi[\sqrt{\beta}, \sqrt{\beta'}]^T, \sqrt{\beta\beta'}\right\}, \left\{[-\sqrt{\beta}, \sqrt{\beta'}]^T, -\sqrt{\beta\beta'}\right\}$ with $\xi \triangleq (\beta + \beta')^{-1/2}$. The spectral radius can then be obtained in closed-form as $r(\boldsymbol{\Lambda}(\boldsymbol{d})) = \sqrt{\beta\beta'}$. Thus the (non-transformed) feasibility set $\mathcal{F}$ is given by all $\widetilde{\boldsymbol{d}} = [\widetilde{d}_1, \widetilde{d}_2]^T$ that satisfies

$$\widetilde{d}_1\widetilde{d}_2 \left(\sum_{j\in\mathcal{J}_1} \frac{g_{2j}}{g_{1j}}\right)\left(\sum_{j\in\mathcal{J}_2} \frac{g_{1j}}{g_{2j}}\right) < 1. \quad (18)$$

Clearly, the feasibility set depends on the channel gains in a non-linear manner. We note that the optimal $(\widetilde{d}_1, \widetilde{d}_2)$ lies on the inner boundary of $\widetilde{\mathcal{F}}$, since to maximize the objective function we must choose $\widetilde{d}_1$ or $\widetilde{d}_2$, or both, to be as large as possible. Moreover, the following observations can be made for the three utility functions.

LIN utility: The transformed feasibility set $\widetilde{\mathcal{F}} = \mathcal{F}$ is unchanged; see Fig. 1(a). We include constraint (17d) which can be written as $y_i \le Y_i, i = 1, 2$; the actual value for Y_i depends on the optimal demand vector for the complementary network. To maximize the sum utility, clearly the optimal solution is to assign either $y_1^\star = Y_1$ or $y_2^\star = Y_2$, i.e., an extreme solution. Moreover, the optimal solution is unique.

LOG utility: The transformed feasibility set $\widetilde{\mathcal{F}}$, including the constraint (17d), is a polytope; see Fig. 1(b). To maximize the objective function $k_1y_1 + k_2y_2$, an optimal solution is given by the boundary extreme solution. This conclusion is similar to the linear utility case, except that the optimal solution is unique only if $k_1 \ne k_2$.

DLOG utility: The transformed feasibility set $\widetilde{\mathcal{F}}$, including the constraint (17d), is strictly convex; see Fig. 1(c). To maximize the objective fuction $k_1y_1 + k_2y_2$, the optimal solution is not necessarily an extreme solution, but is always unique. This suggests the fairest data offloading, as neither of the demands is likely to be very small.

In the next section, we shall use more sophisticated analytical tools to shed further insight on the convexity of the transformed feasibility set $\widetilde{\mathcal{F}}$ for any n.

D. Arbitrary Number of Base Stations

For larger n, the spectral radius cannot be computed in closed-form, and it is expected that the dependence on the channel gains remains non-linear and complicated. Nevertheless, an efficient numerical approach is warranted for arbitrary number of base stations n. Theorem 3 states the convexity of the feasibility set $\widetilde{\mathcal{F}}$ or its complement $\widetilde{\mathcal{F}}^c$.

Theorem 3: The following convexity results hold.

LIN utility: $\widetilde{\mathcal{F}}^c$ is convex for $n = 2$. But $\widetilde{\mathcal{F}}^c$ is generally not convex for $n \geq 3$.

LOG utility: $\widetilde{\mathcal{F}}$ is convex for $n = 2$. Moreover, $\widetilde{\mathcal{F}}$ is strictly convex for $n \geq 3$.

DLOG utility: $\widetilde{\mathcal{F}}$ is strictly convex for any $n \geq 2$.

Proof: The proof for $n = 2$ for all cases was given in Section IV-C. We now consider the case $n \geq 3$ by applying the results in [10], which are closely related to the well-known Perron-Frobenius theorem. First, note that we can express

$$\mathbf{\Lambda}(\boldsymbol{g}(\boldsymbol{y})) = \mathrm{diag}(g(y_1), \cdots, g(y_n))\, \widetilde{\mathbf{\Lambda}} \tag{19}$$

where $g(y_i)$ is the ith element of $\boldsymbol{g}(\boldsymbol{y})$ and the (i,k)th element of $\widetilde{\mathbf{\Lambda}}$ is

$$\widetilde{\lambda}_{ik} = \begin{cases} 0, & \text{if } i = k; \\ \sum_{j \in \mathcal{J}_i} g_{kj}/g_{ij}, & \text{if } i \neq k \end{cases}. \tag{20}$$

LIN utility: We have $g(y) = y$. Applying [10, Theorem 1.60] known as the linear mapping case to (19), we obtain that $\widetilde{\mathcal{F}}^c$ is in general not convex.

LOG utility: We have $g(y) = \exp(y)$. The matrix structure in (19) is referred to as the exponential mapping case in [10]. Moreover, $\widetilde{\mathbf{\Lambda}}$ and $\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{\Lambda}}^T$ are irreducible; see Lemma 3 with definition of irreducibility in the Appendix B. These two conditions allow us to apply [10, Theorem 1.63] to show that $\widetilde{\mathcal{F}}$ is strictly convex for $n \geq 3$

DLOG utility: We have $g(y) = \exp(\exp(y)) - 1$. The following inequality holds after some calculus and algebraic manipulations:

$$\begin{aligned} & d\frac{\partial^2 U(d)}{\partial d^2} + \frac{\partial U(d)}{\partial d} \\ &= \frac{\partial U(d)}{\partial d}\left(1 - \frac{d}{1+d}\left(1 + \frac{1}{\log(1+d)}\right)\right) \\ &< \frac{\partial U(d)}{\partial d}\left(1 - \frac{d}{1+d}\left(1 + \frac{1}{d}\right)\right) = 0 \end{aligned} \tag{21}$$

where the above inequality is due to $\log(1+d) < d$ for $d > 0$. From Lemma 4 in Appendix C with x and $f(x)$ replaced by d and $U(d)$, respectively, the inverse of $U(d)$, i.e., $g(y)$, is strictly log-convex. Since all diagonal elements of $\mathrm{diag}(\boldsymbol{g}(\boldsymbol{y}))$ are strictly log-convex, by [10, Corollary 1.46], it follows that $\widetilde{\mathcal{F}}$ is strictly convex. ∎

The number of users does not significantly affect the complexity of the optimization problem $P1$, due to the same-demand assumption that we have imposed. Instead, the complexity of the optimization problem depends on n, the number of transmitters, e.g., base stations or access points. Assuming LOG or DLOG utility is used, Theorem 3 states that the feasibility set is convex, and hence the complexity for large n is still manageable with the use of convex optimization techniques [13].

Remark 3 (Generalizing Utility Function): Theorem 3 applies to a more general class of utility function $U(d)$. Specifically, if the utility function satisfies $d\frac{\partial^2 U(d)}{\partial d^2} + \frac{\partial U(d)}{\partial d} < 0$ for $n \geq 3$, then $\widetilde{\mathcal{F}}$ is strictly convex. We note that DLOG is a special case, see (21). This conclusion follows immediately from the proof for Theorem 3, in which Lemma 4 was used to show that $g(y)$ is strictly log-convex. Moreover, it follows that $g(y)$ is convex and so the constraint (17d) is convex as its left-hand side is a sum of convex functions. Hence the optimization problem $P1$ is a convex optimization for this general class of utility functions.

E. Algorithm to Limit Maximum Load

So far in our analysis, we consider feasible load $\boldsymbol{x}^\star \geq 0$, which holds if the spectral-radius constraint is strictly less than one. In practice, the load cannot exceed one, due to limited availability of network resources. To impose a constraint $0 \leq \boldsymbol{x}^\star \leq 1$ explicitly in Problem $P0$ however appears challenging. Instead, in this section, we propose an iterative algorithm that reduces the demand such that $0 \leq \boldsymbol{x}^\star \leq 1$.

1) Preliminaries: Let us consider the following optimization problem that is generalized from Problem $P0$. Define Problem $Q(\rho)$, where $0 \leq \rho \leq 1$ is an arbitrary but fixed constant, to be the same as Problem $P0$ but with the spectral-radius constraints (16b) and (16c) replaced by $r(\mathbf{\Lambda}(\boldsymbol{g}(\boldsymbol{y}))) < \rho$ and $r(\mathbf{\Lambda}'(\boldsymbol{g}(\boldsymbol{y}'))) < \rho$, respectively. We denote the corresponding feasibility sets as $\mathcal{F}(\rho)$ and $\mathcal{F}'(\rho)$, respectively. Clearly, Problem $Q(\rho)$ specializes to Problem $P0$ if $\rho = 1$. Corresponding to Problem $Q(\rho)$, the optimal demand vector and load vector are denoted respectively as $\boldsymbol{d}^\star(\rho)$ and $\boldsymbol{x}^\star(\rho)$ for the regular cellular network, and similarly $\boldsymbol{d}'^\star(\rho)$ and $\boldsymbol{x}'^\star(\rho)$ for the WiFi network. Finally, we denote the maximum optimal load as $x^\star_{\max}(\rho) \triangleq \max\{x^\star_i(\rho), x'^\star_j(\rho), i \in \mathcal{N}, j \in \mathcal{N}'\}$. Thus, $\boldsymbol{x}^\star(\rho) \leq 1$ if and only if $x^\star_{\max}(\rho) \leq 1$.

We note that all the analysis so far for Problem $P0$ apply also for Problem $Q(\rho)$, independent of the actual value of ρ. Thus, the numerical solution for Problem $Q(\rho)$ can be obtained similarly as for Problem $P0$. It is useful to note that $\boldsymbol{x}^\star(\rho)$ and $x^\star_{\max}(\rho)$ with $\rho = 1$ correspond to the optimal values for the special case of Problem $P0$.

If $x^\star_{\max}(1) \leq 1$, then $\boldsymbol{x}^\star(1)$ is an optimal solution for Problem $P0$ *and* satisfies the required constraint $0 \leq \boldsymbol{x} \leq 1$. Henceforth, we assume that $x^\star_{\max}(1) > 1$. In the following, we first propose an algorithm such that the final load vector satisfies $0 \leq \boldsymbol{x} \leq 1$, followed by the theoretical justifications.

2) Algorithm: To ensure the load is limited by one, we propose to use the demand vector $\boldsymbol{d}^\star(\rho)$ corresponding to the load vector solution $\boldsymbol{x}^\star(\rho)$ in Problem $Q(\rho)$, where ρ is determined by the solution of the following optimization problem:

$$(P2) \quad \max_{0 \leq \rho < 1} \rho \tag{22}$$

$$\text{s.t. } x^\star_{\max}(\rho) \leq 1. \tag{23}$$

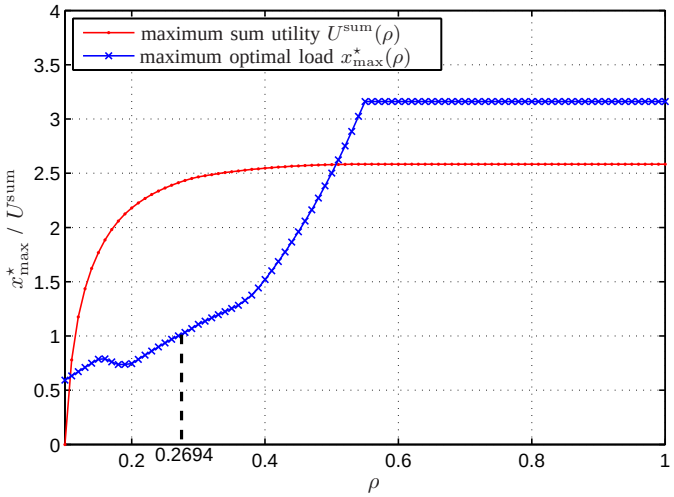


Fig. 2. The graph of maximum optimal load $x^{\star}_{\max}$ and optimal objective value U^{sum} over ρ. The value $\rho^{\star} = 0.2694$ corresponds to $x^{\star}_{\max}(\rho^{\star}) = 1$.

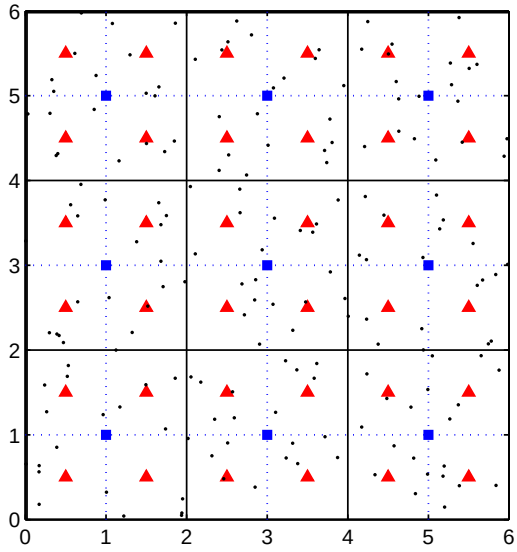


Fig. 3. Network configuration: base stations, access points and users are shown with the blue squares, red triangles, and black dots, respectively.

That is, ρ is the largest possible value such that $0 \leq \boldsymbol{x}^{\star}(\rho) \leq 1$. In general, $x^{\star}_{\max}(\rho)$ is not a monotonic function of ρ. For example, see Fig. 2, where the detailed scenario setup is described in Section V. Nevertheless, since we have reduced the optimization to only one variable, an exhaustive search based on a finely-quantized interval over $0 \leq \rho < 1$ can be performed to solve Problem $P2$, where for each ρ Problem $Q(\rho)$ is solved. This method shall be employed to obtain numerical results in Section V.

3) Theoretical Basis: The theoretical basis for the above algorithm stems from Theorem 4 and Theorem 5 below. Theorem 4 ensures that the highest possible sum utility is achieved for Problem $Q(\rho)$ if we choose ρ to be as large as possible. Theorem 5 ensures the existence of a solution in Problem $P2$ under the equal-demand assumption.

Theorem 4: Denote the optimal sum utility value for Problem $Q(\rho)$ as $U^{\text{sum}}(\rho), 0 \leq \rho \leq 1$. Then $U^{\text{sum}}(\rho)$ is a strictly increasing function of ρ.

Proof: Let $\widetilde{\boldsymbol{d}}' = \widetilde{\boldsymbol{d}} + \boldsymbol{e}, \boldsymbol{e} \geq 0$. It can be easily checked from definition (9) that $\boldsymbol{E} \triangleq \boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}') - \boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}) \geq 0$, with equality if and only if $\boldsymbol{e} = 0$. Similar to the proof that $\widetilde{\boldsymbol{\Lambda}}$ is irreducible in Lemma 3, it can be shown that $\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}) \geq 0$ is irreducible. Thus, we can apply Lemma 5 in Appendix D to get $r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}')) = r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}) + \boldsymbol{E}) \geq r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}))$, with equality if and only if $\boldsymbol{e} = 0$. Thus, $\widetilde{\boldsymbol{d}}' \geq \widetilde{\boldsymbol{d}}$ if and only if $r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}')) \geq r(\boldsymbol{\Lambda}(\widetilde{\boldsymbol{d}}))$. This implies that the feasibility set $\mathcal{F}(\rho)$ (and similarly for $\mathcal{F}'(\rho)$) satisfies $\mathcal{F}(\rho_1) \subset \mathcal{F}(\rho_2)$ for $\rho_1 < \rho_2$. Thus, $U^{\text{sum}}(\rho_1) < U^{\text{sum}}(\rho_2)$ for $\rho_1 < \rho_2$, i.e., $U^{\text{sum}}(\rho)$ is a strictly increasing function. ■

For illustration, we plot the sum utility $U^{\text{sum}}(\rho)$ (added by a constant such that it becomes positive) as a function of ρ in Fig. 2. In accordance with Theorem 4, the sum utility is increasing with ρ.

Theorem 5: Consider Problem $Q(\rho)$ where $x^{\star}_{\max}(\rho) > 1$ for $\rho = 1$. Then there exists an optimal load vector $\boldsymbol{x}^{\star}(\rho)$ such that $x^{\star}_{\max}(\rho) = 1$ for some $0 < \rho < 1$.

Proof: From the proof of Theorem 4, the feasible set $\widetilde{\mathcal{F}}(\rho)$ becomes strictly smaller as ρ decreases. From Theorem 2, the load vector is a monotonic function of the demand vector. Thus, every load vector corresponding to a demand vector in the feasible set also decreases in value, as ρ decreases. For sufficiently small $\rho \to 0$, all the elements of the optimal demand must approach the all-zero vector and thus $x^{\star}_{\max} \to 0$. By continuity, there exists $x^{\star}_{\max} = 1$ for some $0 < \rho < 1$. ■

For illustration, we plot the maximum optimal load $\boldsymbol{x}^{\star}(\rho)$ as a function of ρ in Fig. 2. We note that, in contrast to $U^{\text{sum}}(\rho)$, $\boldsymbol{x}^{\star}(\rho)$ is not necessarily an increasing function of ρ. Nevertheless, there exists $x^{\star}_{\max}(\rho) = 1$ as ρ is decreased from $\rho = 1$, in accordance with Theorem 5. From Fig. 2, we see that the largest $0 \leq \rho < 1$ such that $x^{\star}_{\max}(\rho) = 1$ is given by $\rho = \rho^{\star} = 0.2694$. Thus, this gives the solution for Problem $P2$. The corresponding demand and load allocation are shown as Fig. 5 later in Section V.

V. Numerical Results

In this section, unless otherwise specified, we obtain numerical results assuming the utility function is the LOG utility. The optimization problem $P1$, and more generally $Q(\rho)$, is convex. This is because the objective function is linear and the constraint set is convex due to Theorem 3 for the case of LOG utility. Thus, the optimal demand vectors $\widetilde{\boldsymbol{d}}^{\star}, \widetilde{\boldsymbol{d}}'^{\star}$ can be solved efficiently by standard numerical solvers. Specifically, we use the active-set algorithm with the fmincon function in the MATLAB software. The optimal load vectors $\boldsymbol{x}^{\star}$ and $\boldsymbol{x}'^{\star}$ are then computed using a synchronous or asynchronous iterative algorithm according to Lemma 1 and Remark 1, respectively.

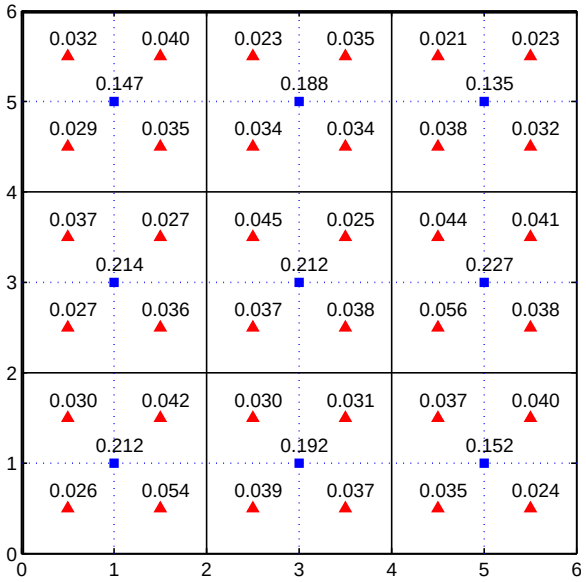
Our theoretical result and numerical approach apply regardless of where the base stations, access points and users are deployed. For ease of viewing the numerical results, we position the base stations and access points at equal distance apart, while the users are at arbitrarily but fixed locations (obtained by the realizations from a uniform distribution). For

the network configuration, we assume all cells are square in shape. The regular cellular network consists of $n = 9$ cells, where each square cell is of two unit length. The cells are arranged uniformly as shown in Fig. 3. A base station is placed in the centre of each regular cell, shown as a blue square in Fig. 3. In each regular cell, there are four disjoint square WiFi cells each of unit length, making a total of $n' = 36$ WiFi cells. Within each WiFi cell, there are 5 users. A WiFi access point is placed in the center of each WiFi cell, shown as a red triangle. Every user is served by the WiFi cell and the base station cell that it resides in. Thus, every access point can support up to 5 users while every base station can support up to 20 users. We make the same-demand assumption that all users in the same (regular or WiFi) cell are allocated the same demand.

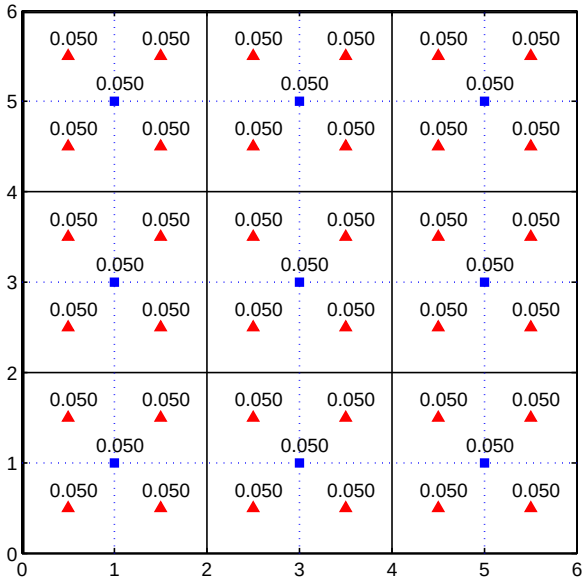
We use the same weight $k_{ij} = 1$ for all base stations, and the weight $k'_{ab} = 1/4$ for all access points; the difference in weights is used to account for the fact that the number of access points is four times the number of base stations. We set the (normalized) transmission power of every regular cell as 100, the transmission power of every WiFi cell as 1, and the noise variance as 0.01. The channel gain from the ith regular cell to the jth user is fixed as $g_{ij} = z_{ij}^{-\kappa}$ where z_{ij} is the distance between transmitter i and receiver j, and $\kappa = 4$ is the path loss exponent. The channel gains for the WiFi cells are obtained similarly.

A. Low Maximum Demand

In our first numerical experiment, we fix the maximum demand to be $D_i = 0.1, i \in \mathcal{N}$. Solving for Problem $P1$ numerically, we obtain the optimal demand and load allocations as indicated in Fig. 4 besides the positions of the base stations and access points. From Fig. 4(a), all cells are operating below full load, i.e., $x^\star_{\max} \leq 1$. We observe that the load allocation in Fig. 4(a) is non-uniform, due to the non-uniform user distribution as shown in Fig. 3. From Fig. 4(b), the optimal demand to be served by every regular cell and WiFi cell is the same, given by $d^\star = 0.05$ nat. Thus, all users are served the maximum rate of $D_i = 0.1$ nat in total, with $d^\star$ contributed by the regular cell and another $d^\star$ contributed equally by the WiFi cell. The reason for such a uniform distribution of the optimal demand follows. We observe that the optimal demand is also given by $d^\star = 0.05$ if we maximize the sum utility without the spectral-radius constraints (17b) and (17c) (not shown here). This implies that the spectral-radius constraints are in fact not active in the original problem, i.e., the demands can be treated as unconstrained variables without loss of optimality. For our choice of $k_{ij} = 1, k'_{ab} = 1/4$ and with one base station for every four access points, the optimal demand is thus uniform for all the access points and the base stations. In further numerical experiments where the weights k_{ij}, k'_{ab} are changed (not shown here), we observe that the optimal demand is not necessarily uniform, i.e., the values are different for the base stations and the access points. Nevertheless, we make the consistent observation that the same optimal demand is obtained whether with or without the spectral-radius constraints; this is consistent with the earlier observation that the maximum load has not exceeded one.



(a) Optimal load allocation.



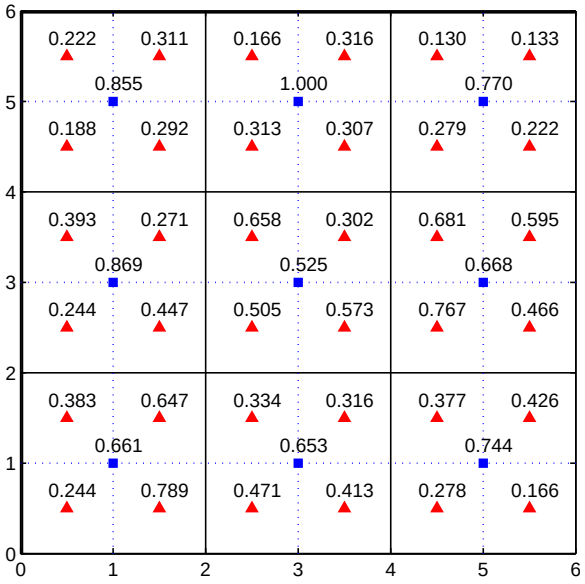
(b) Optimal demand allocation.

Fig. 4. Optimal allocation with LOG utility and maximum demand fixed as 0.1. The spectral radius is subject to a maximum constraint of 1; the maximum load value turns out to be less than one.

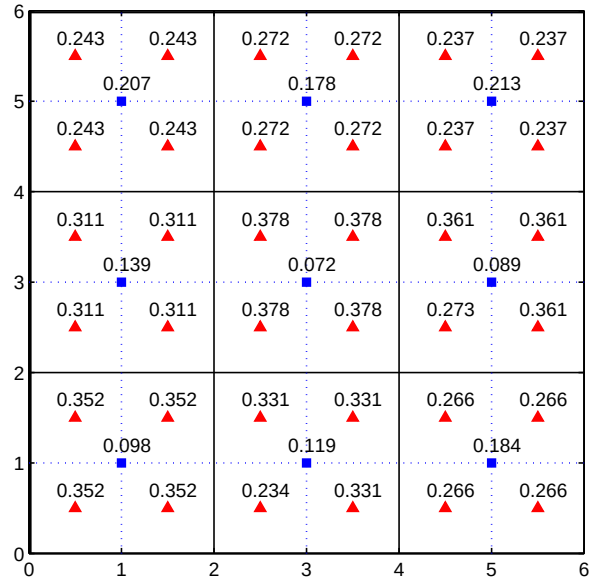
B. High Maximum Demand

Next, we increase the maximum demand to $D_i = 0.45, i \in \mathcal{N}$. With this high maximum demand, we shall see that the spectral-radius constraint becomes active, and the maximum demand requested by the users cannot be achieved.

Solving for Problem $P1$ numerically, we obtain the maximum optimal load as $x^\star_{\max}(\rho) > 1$ with $\rho = 1$. Thus, some of the cells are overloaded and the optimal demand vector $\boldsymbol{d}^\star(\rho)$ with $\rho = 1$ cannot be practically implemented. To reduce the load, we solve Problem $P2$ via the algorithm proposed in Section IV-E. In this algorithm, we obtain $\rho^\star$ given by the largest ρ in Problem $Q(\rho)$ such that the corresponding maximum optimal load $x^\star_{\max}(\rho) = 1 - \epsilon$ where $\epsilon > 0$ is close to zero. The sum utility U^{sum} and the maximum optimal load $x^\star_{\max}$ are plotted as functions of ρ in Fig. 2. For ease of



(a) Optimal load allocation.



(b) Optimal demand allocation.

Fig. 5. Optimal allocation with LOG utility and maximum demand fixed as 0.45. The spectral radius is subject to a maximum constraint of 0.2694 so that the maximum load value is less than one.

viewing, U^{sum} has been increased by a constant value. From Fig. 2, the optimal ρ is $\rho^\star = 0.2694$. The average per-user demand achieved, obtained by averaging the sum demand over all users, is then given by 0.4449 nat, which is less than the maximum demand of $D_i = 0.45$.

After constraining the load such that $x^\star_{\max}(\rho^\star) \leq 1$ via Problem $P2$, we obtain the optimal demand vector $\boldsymbol{d}(\rho^\star)$, and the corresponding load $\boldsymbol{x}^\star(\rho^\star)$, as shown in Fig. 5. From Fig. 5(a), all the loads have been constrained to less than one. Similar to the low maximum demand case, the load allocation is not uniform due to the non-uniform user allocation. However, in contrast to the uniform demand allocation for the low maximum demand case shown in Fig. 4(b), the demand allocation in Fig. 5(b) is not uniform. For example, in Fig. 5(b), the base station at coordinates $(5, 3)$ serves 0.089 nat to all its users, while the WiFi access points within the base station cell serve a variation of demand, ranging from 0.273 nat for the access point at $(4.5, 2.5)$, to 0.361 nat for the access points at $(4.5, 3.5)$, $(5.5, 2.5)$ and $(5.5, 3.5)$. That is, the users are served in total a demand ranging from 0.362 nat to the maximum request demand of 0.45 nat. The reason in the difference of the demand served is likely because the users that are served a smaller demand are closer to the center of the entire network and hence received more interference. We note that this observation may not always hold in general since it depends on the user distribution and the resulting optimal load allocation in the entire network. For example, all the users served by the base station cell at $(3, 3)$ receive the maximum demand, with 0.072 nat from the base station and 0.378 nat from their respective access points. In general, however, we may still conclude that for the high maximum demand case, the spectral-radius constraints become tight which can limit the demand served to the users, especially those cells that receive the most amount of interference.

C. Different Utility Functions

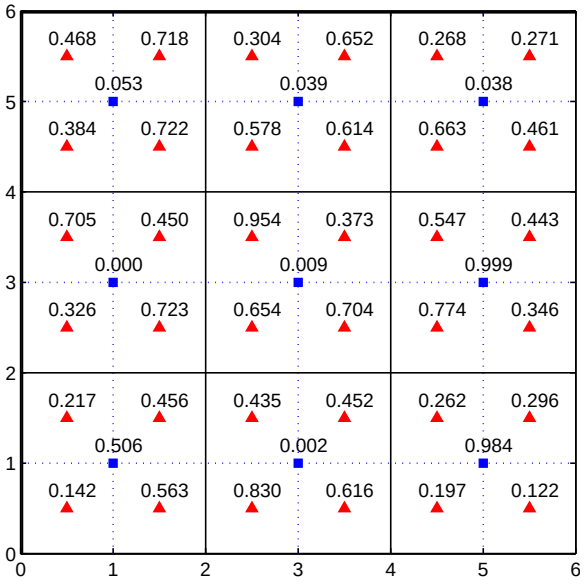
We assume the high-maximum-demand case of $D_i = 0.45, i \in \mathcal{N}$. The results for the LOG utility has been given earlier in Fig. 5. The results for DLOG utility are almost identical to the case of the LOG utility, and are thus omitted. The similarity of the result is likely because the user fairness has already been largely taken into account via the LOG utility, and emphasizing this same aspect via the DLOG utility does not lead to a significantly different optimal solution.

Finally, we consider the use of the LIN utility. From Theorem 3, the feasible set $\widetilde{\mathcal{F}}$ may not be convex and hence our numerical solution is not necessarily optimal. Nevertheless, we shall see that we can still obtain a higher average per-user demand, but at the expense of user fairness.

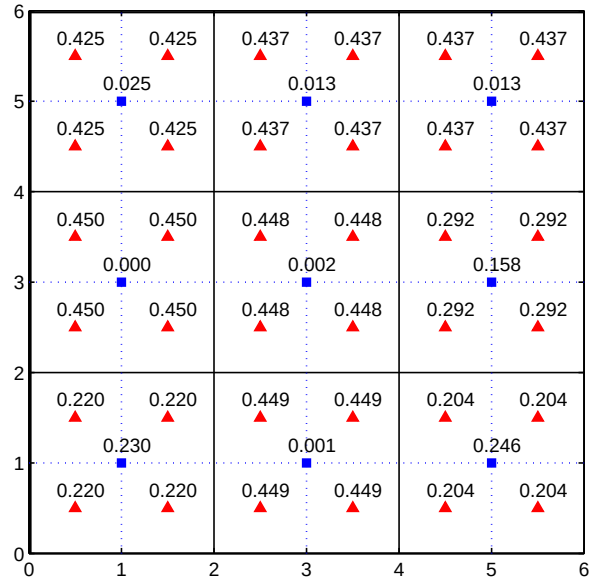
The optimal demand vector $\boldsymbol{d}(\rho^\star)$, and the corresponding load $\boldsymbol{x}^\star(\rho^\star)$, are shown in Fig. 6. Again, we have constrained the load to be less than one, similarly by solving Problem $P2$ as before. The average per-user demand achieved is observed to be the maximally possible given by 0.45 nat, compared to 0.4449 nat achieved with LOG utility. This is within expectation since the LIN utility only focus on maximizing the sum demand. However, not all base stations or access points are uniformly served similar demand. In extreme cases, it is possible that some users are not served at all while other users are served the maximum load.

VI. Conclusion

We have presented a utility-based optimization framework for data offloading in cellular networks, taking into account the inherent coupling relation among the cells. Within this framework, fundamental properties on the computation, feasibility, and monotonicity of the load-coupled system have been studied. Three utility functions that differ in the emphasis on fairness have been considered, and fundamental insights of convexity analysis of the resulting optimization problem have been developed. Our analysis shows that optimal offloading is tractable when fairness is stressed. We also propose a strategy to constrain the load to some maximum value, as required for



(a) Optimal load allocation.



(b) Optimal demand allocation.

Fig. 6. Optimal allocation with LIN utility and maximum demand fixed as 0.45. The spectral radius is subject to a maximum constraint of 0.301 so that the maximum load value is still less than one.

practical implementation, and provide theoretical justification for the proposed algorithm. In conclusion, our work provides a structured view on the offloading problem, and our analysis serves as a theoretical reference for empirical simulations and further performance evaluation. As future work, we shall consider the related problems of energy minimization and user-network association.

Appendix A
Proof of Theorem 2

Consider the following asynchronous iteration in Remark 1 that runs for $k = 1, \cdots, K$. For each *outer iteration* k, we execute an *inner iteration* that runs for $m = 1, \cdots, n$:

$$x_m^k = f_m(\boldsymbol{x}^{k-1,m}) \tag{24}$$

where $\boldsymbol{x}^{0,1} = [x_1^0, \cdots, x_n^0]$ is an arbitrary initial load and $\boldsymbol{x}^{k-1,m} \triangleq [x_1^k, \cdots, x_{m-1}^k, x_m^{k-1}, \cdots, x_n^{k-1}]$ denotes the most current updated load vector. After all iterations, the final load vector is given by $\boldsymbol{x}^{K,n+1}$, denoted simply as $\boldsymbol{x}^K$. From Remark 1, (24) is an asynchronous iteration which ensures that $\boldsymbol{x}^K$ converges to the final fixed-point solution in (5).

Suppose only one element of $\boldsymbol{d}'$ is strictly greater than $\boldsymbol{d}$, say d_{ij}. For the initial load, we choose $\boldsymbol{x}^{1,0} = \boldsymbol{x}^\star$. We then obtain the following results by performing the iteration (24).

- For $k = 1$: $x_i^1 > x_i^0$ while $x_\ell^1 = x_\ell^0$ for $\ell \neq i$.
- For $k = 2$: $x_i^2 = x_i^1$ while $x_\ell^2 > x_\ell^1$ for $\ell \neq i$.
- For $k \geq 3$: $x_i^k > x_i^{k-1}$ while $x_\ell^k > x_\ell^{k-1}$ for $\ell \neq i$.

Specifically, for $k = 1$, we have used (4) where we replace d_{ij} by d'_{ij}; for $k \geq 2$, we have used (24) and the result for the prior k. Thus, $\boldsymbol{x}^{k,n} > \boldsymbol{x}^{k-1,n}$ for $k \geq 3$, while $\boldsymbol{x}^{2,n} \geq \boldsymbol{x}^{1,n} \geq \boldsymbol{x}^{1,0} = \boldsymbol{x}^\star$. Together with the convergence guarantee, we get $\lim_{K\to\infty} \boldsymbol{x}_n^K = \boldsymbol{x}'^\star > \boldsymbol{x}^\star$ as desired. It is easy to check that the above conclusion holds even if more than one element in $\boldsymbol{d}'$ is strictly greater than $\boldsymbol{d}$ if all the users are served by the same (and only) base station i.

Next, consider the general case where $\boldsymbol{d}' \geq \boldsymbol{d}, \boldsymbol{d}' \neq \boldsymbol{d}$. Let $s \geq 1$ be the number of base stations serving users with different demand in $\boldsymbol{d}'$ and $\boldsymbol{d}$. Then we can always find a set $\{\widetilde{\boldsymbol{d}}_1, \cdots, \widetilde{\boldsymbol{d}}_s\}$ with distinct elements ordered according to $\boldsymbol{d}' \geq \widetilde{\boldsymbol{d}}_s \geq \cdots \geq \widetilde{\boldsymbol{d}}_1 \geq \boldsymbol{d}$ such that for any neighbouring pairs of vectors, e.g. $\{\boldsymbol{d}', \widetilde{\boldsymbol{d}}_s\}$, only one base station serve the users with different demand. We then use the following inductive steps to complete the proof. First, we obtain the load $\widetilde{\boldsymbol{x}}_1$ that corresponds to $\widetilde{\boldsymbol{d}}_1$ in (5). To do so, we use $\boldsymbol{x}^\star$ as the initial load and the asynchronous iteration as before, which shows that $\widetilde{\boldsymbol{x}}_1 > \boldsymbol{x}^\star$. Second, we use $\widetilde{\boldsymbol{x}}_1$ as the initial load and the asynchronous iteration as before, to show that the corresponding load $\widetilde{\boldsymbol{x}}_2$ satisfies $\widetilde{\boldsymbol{x}}_2 > \widetilde{\boldsymbol{x}}_1$. Proceeding similarly, we thus get $\boldsymbol{x}'^\star > \widetilde{\boldsymbol{x}}_s > \cdots > \widetilde{\boldsymbol{x}}_1 > \boldsymbol{x}^\star$.

Appendix B
Lemma on Irreducible Matrix

Consider a non-negative matrix $\boldsymbol{B} \in \mathcal{R}_+^{n\times n}$ with the (i,j)th element given by b_{ij}. Let the *incidence matrix* of $\boldsymbol{B}$ be $\boldsymbol{A} \in \{0,1\}^{n\times n}$ with the (i,j)th element $a_{ij} = 1$ if $b_{ij} > 0$ and $a_{ij} = 0$ if $b_{ij} = 0$. Denote the element of $\boldsymbol{A}^m$ as $a_{ij}^{(m)}$. We say $\boldsymbol{B}$ is *irreducible* if $a_{ij}^{(m)} > 0$ for all i, j for some $m \geq 1$.

Lemma 3: The matrix $\widetilde{\boldsymbol{\Lambda}} \in \mathcal{R}_+^{n\times n}, n \geq 3$, with (i,k)th element given by (20) is irreducible. Also, $(\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^T)$ is irreducible.

Proof: From (20), the diagonal elements of $\widetilde{\boldsymbol{\Lambda}}$ are zeros, while the off-diagonal elements are strictly positive since the channel gains $\{g_{kj}\}$ are positive. Thus, the incidence matrix of $\widetilde{\boldsymbol{\Lambda}}$ is $\boldsymbol{A} = \mathbf{1}_n \cdot \mathbf{1}_n^T - \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the n-by-n identity matrix. Thus $\boldsymbol{A}^2 = (n-2)\mathbf{1}_n \cdot \mathbf{1}_n^T + \boldsymbol{I}_n$. Clearly $\boldsymbol{A}^2 > 0$, and so $\boldsymbol{\Lambda}(\boldsymbol{d})$ is irreducible. Moreover, $(\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^T)$ is irreducible as the incidence matrix is $\boldsymbol{A}^2 > 0$. ■

Appendix C
Lemma on Log-Convexity to prove Theorem 3

Lemma 4: Assume $f(x)$ is an increasing function with inverse $g(y) = f^{-1}(y)$. Assume $f(x)$ and $g(y)$ are differentiable. Then $g(y)$ is strictly log-convex if and only if

$$xf''(x) + f'(x) < 0. \tag{25}$$

where $f'(\cdot)$ and $f''(\cdot)$ are the first and second derivatives of $f(\cdot)$.

Proof: The second derivative of $\log(g(y))$ is given by $g''(y)/g(y) - (g'(y)/g(y))^2$, so $g(y)$ is strictly log-convex iff

$$g''(y)g(y) - (g'(y))^2 > 0. \tag{26}$$

To complete the proof, we shall show that (26) holds.

We can write $g(f(x)) = x$. Differentiating with respect to x, we get $g'(f(x)) = 1/f'(x)$. Differentiating again with respect to x, we get $g''(f(x)) = -f''(x)/(f'(x))^3$. Thus the left-hand side of (26) can be written as $g''(f(x))g(f(x)) - (g'(f(x)))^2 = -xf''(x)/(f'(x))^3 - 1/(f'(x))^2 = -(xf''(x) + f'(x))/(f'(x))^3$. Since $f'(x) \geq 0$, (26) holds if and only if (25) holds, which completes the proof. ■

Appendix D
Lemma to Prove Theorem 4

Lemma 5: Let $\boldsymbol{A}, \boldsymbol{B} \geq 0$ be n-by-n matrices, and $\boldsymbol{A}$ is irreducible. Then $r(\boldsymbol{A}+\boldsymbol{B}) \geq r(\boldsymbol{A})$ with equality if and only if $\mathbf{B} = \mathbf{0}$.

Proof: Let $\boldsymbol{u}$ be the (right) eigenvector that corresponds to the largest eigenvalue of $\boldsymbol{A}$. Applying the Perron-Frobenius Theorem to $\boldsymbol{A}$ [10], we have $\boldsymbol{u} > 0$ and $r(\boldsymbol{A})$ equals the largest eigenvalue. Then the spectral radius of $\boldsymbol{A}+\boldsymbol{B}$ can be written as

$$r(\boldsymbol{A}+\boldsymbol{B}) = \max_{\|\boldsymbol{z}\|=1} |\boldsymbol{z}^H(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{z}| \tag{27}$$

$$\geq \boldsymbol{u}^H\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}^H\boldsymbol{B}\boldsymbol{u} \tag{28}$$

$$\geq \boldsymbol{u}^H\boldsymbol{A}\boldsymbol{u} = r(\boldsymbol{A}) \tag{29}$$

Here, the first inequality is due to replacing $\boldsymbol{z}$ by $\boldsymbol{u}$. The second inequality is due to $\boldsymbol{B} \geq 0$ and $\boldsymbol{u} > 0$, and becomes an equality if and only if $\mathbf{B} = \mathbf{0}$. This concludes the proof. ■

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016," Feb. 2012, White Paper. [Online]. Available: http://tinyurl.com/b9berc

[2] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, Apr. 2013.

[3] B. Han, P. Hui, and A. Srinivasan, "Mobile data offloading in metropolitan area networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 14, no. 4, pp. 28–30, Oct. 2011.

[4] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," *IEEE Netw.*, vol. 21, no. 1, pp. 27–33, Jan./Feb. 2007.

[5] I. Siomina, A. Furuskar, and G. Fodor, "A mathematical framework for statistical QoS and capacity studies in OFDM networks," in *Proc. Personal, Indoor Mobile Radio Commun. Symp.*, Sept. 2009, pp. 2772 –2776.

[6] K. Majewski and M. Koonert, "Conservative cell load approximation for radio networks with Shannon channels and its application to LTE network planning," in *Proc. Sixth Advanced Int. Conf. Telecommun. (AICT)*, May 2010, pp. 219–225.

[7] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[8] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for interference-coupled cellular data networks," in *Proc. IEEE Int. Conf. Commun.*, June 2012, pp. 5102–5107.

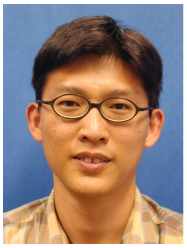
[9] C. K. Ho, D. Yuan, and S. Sun, "Data offloading in load coupled networks: Solution characterization and convexity," in *Proc. IEEE Int. Conf. Commun.*, June 2013, pp. 5102–5107.

[10] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of Resource Allocation in Wireless Networks*, ser. Foundations in Signal Processing, Commun. and Networking. Berlin, Germany: Springer, 2009, vol. 3.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[12] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sept. 1995.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.



Chin Keong Ho received the B. Eng. (First-Class Hons., Minor in Business Admin.) and M. Eng. degrees from the Department of Electrical Engineering, National University of Singapore in 1999 and 2001, respectively. He received the Ph.D. degree at the Eindhoven University of Technology, The Netherlands, where he concurrently conducted research work in Philips Research.

Since August 2000, he has been with Institute for Infocomm Research (I^2R), A*STAR, Singapore. He is Lab Head of Energy-Aware Communications Lab, Department of Modulation and Coding. His research interest includes green wireless communications with energy harvesting constraints; cooperative and adaptive wireless communications; and implementation aspects of multicarrier and multiantenna communications.

Di Yuan received his MSc degree in computer science and engineering, and PhD degree in operations research at Linköping Institute of Technology in 1996 and 2001, respectively. At present he is full professor in telecommunications at the Department of Science and Technology, Linköping University, and head of a research group in mobile telecommunications. His current research mainly addresses network optimization of 4G systems. Dr Yuan has been guest professor at Technical University of Milan (Politecnico di Milano), Italy, in 2008, and senior visiting scientist at Ranplan Wireless Network Design Ltd, United Kingdom, in 2009 and 2012. In 2011 and 2013 he has been part time with Ericsson Research, Sweden. He is an area editor of the Computer Networks journal. He has been in the management committee of four European Cooperation in field of Scientific and Technical Research (COST) actions, invited lecturer of European Network of Excellence EuroNF, and Principal Investigator of five European FP7 Marie Curie projects. He is co-recipient of IEEE ICC'12 Best Paper Award.

Sumei Sun (SM'12) received the B.Sc. (with honors) degree from Peking University, Beijing, China; the M.Eng. degree from Nanyang Technological University, Singapore; and the Ph.D. degree from National University of Singapore, Singapore. She has been with Institute for Infocomm Research (I^2R), Agency for Science, Technology, and Research (A*STAR), Singapore, since 1995, where she is currently Head of the Modulation and Coding Department, developing energy- and spectrum-efficient technologies for the next-generation communication systems. Her recent research interests include 5G transmission technologies, renewable energy management and cooperation in wireless systems and networks, and wireless transceiver design. Dr. Sun served as the Technical Program Committee Chair of the 12th IEEE International Conference on Communications Systems (ICCS) in 2010, General Co-chair of the Seventh and Eighth IEEE Vehicular Technology Society Asia Pacific Wireless Communications Symposium, and Track Co-chair of Transmission Technologies of the 75th IEEE Vehicular Technology Conference in 2012. She is also an Associate Editor for IEEE Transactions on Vehicular Technology and an Editor for IEEE Wireless Communication Letters. She was a co-recipient of the 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications Best Paper Award.