Robust User Association for Ultra Dense Networks

Nikolaos Liakopoulos[1,2], Georgios Paschos[1], Thrasyvoulos Spyropoulos[2]

[1]Mathematical and Algorithmic Sciences Lab, FRC, Huawei Technologies SASU, email: firstname.lastname@huawei.com
[2]EURECOM, Sophia-Antipolis France, email: spyropou@eurecom.fr

***Abstract*—We study the user association problem in the context of dense networks, where standard adaptive algorithms become ineffective. The paper proposes a novel data-driven technique leveraging the theory of robust optimization. The main idea is to predict future traffic fluctuations, and use the predictions to design association maps before the actual arrival of traffic. Although the actual playout of the map is random due to prediction error, the maps are robustly designed to handle uncertainty, preventing constraint violations, and maximizing the expectation of a convex utility function, which allows to accurately balance base station loads. We propose a generic iterative algorithm, referred to as GRMA, which is shown to converge to the optimal robust map. The optimal maps have the intriguing property that they jointly optimize the predicted load and the variance of the prediction error. We validate our robust maps in Milano-area traces, with dense coverage and find that we can reduce violations from 25% (achieved by an adaptive algorithm) down to almost zero.**

I. Introduction

The explosion of wireless traffic is driving network operators to deploy heterogeneous sites and constantly increase base station (BS) density, in an effort to improve the spectrum reuse [1]. This trend is expected to culminate with the emerging *Ultra Dense Networks (UDNs)* in the 5G and beyond era, where a user located in an urban area will be surrounded by hundreds of sites, while the available cells may be more than the number of active users [2], [3]. Nevertheless, providing a copious amount of resources is only the first step. A second equally important step is to develop efficient resource management mechanisms in order to balance BS loads, under the high spatio-temporal traffic variability resulting from the small number of users per BS [4]. In the demanding environment of dense 5G networks, user association (choosing a BS for each user among a large number of candidates) and traffic steering (serving a traffic flow from the right BS, carrier, etc.) must be surgically engineered for proper exploitation of the increased density [4].

A number of recent works attempt to formalize the problem and find an optimal solution [4]–[8]. Nevertheless, these frameworks are not ideal for Ultra-Dense Networks for two main reasons:

- *Spatio-temporal variability*: Due to smaller user/site ratios, the traffic demand will vary significantly more over time and space, giving rise to unpredictable traffic spikes.
- *Increased QoS requirements*: With the rise of vertical applications, 5G networks are expected to support slices that provide guaranteed Quality of Service (QoS). Unexpected traffic spikes combined with dynamic association decisions reacting to them, might lead the system to oscillations, instability, and violation of QoS requirements.

Towards addressing these issues, this paper proposes a radically different approach based on two main components:

Data-driven user association maps pre-calculated for each day and time period (e.g. per hour), based on estimates of average traffic demand for that time, day, and location; these maps *proactively* associate users/flows from certain locations to certain BSs, rather than constantly oscillating association decisions due to traffic spikes;

A robust optimization framework for user association that takes into account the prediction error and protects system QoS from traffic spikes. To our best knowledge, this is the first work to propose such robust pre-calculated user association maps for dense heterogeneous networks. Moreover, our work is the first to explore an interesting new user association tradeoff between facilitating traffic prediction vs. improving network performance. These two goals are not always aligned, as it will become clear in our analysis.

A. Related Work and our Contribution

Selecting the association rule vector $\boldsymbol{\pi}$, that assigns each new user to a base station, can be formulated as an optimization problem that targets to maximize a utility function of the resulting base station loads [5]–[8]. Such optimization problems are usually difficult to solve, due to the coupling of the user association decisions; adding a user alters the base station load and affects the performance for all the users connected to it. In the context of UDNs, the size of the problem grows (100s of BSs in small areas with many locations or users to be associated) and mounts an extra difficulty. Complex optimization approaches fall short, since practical systems require fast and lightweight solutions.

The integral user association problem is a combinatorial problem [6], but in [7] the authors show that the solution of their convex relaxation problem is indeed integral, enabling in this way optimization of a general class of objective functions. Later, [8] introduced backhaul constraints in the model. A dynamic biasing scheme is proposed in [5] to load balance heterogeneous wireless cells. All these algorithms and the majority of prior work, assume that traffic characteristics are known at a fine spatiotemporal granularity, which as explained above is a key challenge in UDNs. Although not explicitly addressed in the literature, the above schemes are adaptive and can be used as heuristics in a scenario with unpredictably fluctuating traffic, but it is clear that they can not guarantee satisfaction of QoS requirements in this case. The goal of this paper is to fill this gap in the literature.

A modern trend in networking problems is to tap into the power of available data to deal with uncertainty [9]. For

wireless traffic, many prior works identify structure, like the diurnal pattern during the day or the similarity of traffic during the weekdays and weekends/holidays [10], [11]. However, up to now it is far from clear how to best utilize the data and the observed patterns for improving the performance of user association. *Our contributions in this paper are the following:*

- We propose the idea of precomputing maps, that can be used later to determine user association in real-time.
- To study the map performance we propose an analytical data-driven framework for user association in the context of dense wireless networks with unpredictable traffic spikes. This leads us to the formulation of the robust user association map (RUAM) problem.
- We then propose an efficient generic algorithm (GRMA) that provably produces the optimal robust maps, for a large class of objective functions.
- Finally we demonstrate the efficiency of our framework on real data [12], compared to an adaptive version of a popular user association algorithm [7]. Our simulations show that we achieve more stability (up to 25% improvement) and decreased average latency, especially in periods of high traffic activity. Our framework can adapt to design choices, balancing trade-offs in stability and cost, by tuning the SLA guarantees.

II. Architecture

A. System Model

We consider a region $\mathcal{L} \subseteq \boldsymbol{R}^2$ with ultra dense cellular coverage from a set of $\mathcal{B}$ (possibly heterogeneous) base stations. This region we envision it as a 2D representation of a dense urban environment with fixed base station positions.

Spatial traffic. Users at location $x \in \mathcal{L}$, generate flow requests according to an inhomogeneous Poisson point process with spatial intensity $\lambda(x)$ and have independently and generically distributed file sizes with mean $\frac{1}{\mu}$.

Service Rate. The flows generated at a point $x \in \mathcal{L}$ that are associated to a base station $i \in \mathcal{B}$ are served with rate $C_i(x)$. In our paper, $C_i(x)$ is a location-dependent metric that depicts the wireless signal degradation due to distance of x from the base station $i \in \mathcal{B}$.

$$C_i(x) = W \log(1 + \text{SINR}_i(x)), \tag{1}$$

where W is the available frequency bandwidth, and $\text{SINR}_i(x)$ is given by:

$$\text{SINR}_i(x) = \frac{P_i G_i(x)}{\sum_{j \neq i} P_j G_j(x) + N_0}. \tag{2}$$

P_i denotes the transmission power of base station i, N_0 denotes noise power and $G_i(x)$ is the path loss between the antenna and the UE. This model has been shown to accurately capture the average behavior of wireless systems including shadowing, interference, and path-loss [5], [7], [8], [13].

Association Rules. Let $\pi_i(x) \in [0, 1]$ be the association rules, indicating the fraction of traffic of location x associated to base station i. To associate the total traffic of location x we enforce the constraint $\sum_{i \in \mathcal{B}} \pi_i(x) = 1$. The association variables $\pi_i(x),\ \forall x \in \mathcal{L}$ will be the means to control the performance of the system.

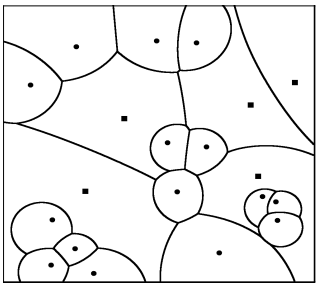


Fig. 1. Representation of an association map.

Base Station Load. The load ρ_i is the fraction of time base station i is busy. The load contribution from a specific location depends on the association rules, and it is equal to $\frac{\lambda(x)}{\mu C_i(x)} \pi_i(x)$. Therefore, considering the area $\mathcal{L}$:

$$\rho_i = \int_{\mathcal{L}} \frac{\lambda(x)}{\mu C_i(x)} \pi_i(x) dx. \tag{3}$$

The vector of base station loads $\boldsymbol{\rho} = (\rho_i)$ is an important performance metric of the system. For example, [14] suggests that assuming a temporal fair scheduler (e.g. round robin, proportional fair) the dynamics of the base station queues can be accurately described by an M/G/1 processor sharing system, where the expected number of active users at base station i is given by $\mathbb{E}[N_i] = \frac{\rho_i}{1-\rho_i}$. This is tightly related with average response time for a flow in base station i, which from Little's law is $\mathbb{E}[T_i] = \frac{1}{\lambda_i} \frac{\rho_i}{1-\rho_i}$, and with the average delay experienced at a location $x \in \mathcal{L}$: $\mathbb{E}[T|X = x] = \sum_{i \in \mathcal{B}} \frac{1}{\mu C_i(x)(1-\rho_i)} \pi_i(x)$, which is derived from the flow throughput equation in [14].

In the following sections we will focus on how to choose association rules $\pi_i(x), \forall x \in \mathcal{L}$ to achieve specific vectors $\boldsymbol{\rho}$ that correspond to important network-wide objectives, e.g. total throughput, average queuing delay, or balancing base station loads.

B. User Association Maps

In this paper we are interested in precalculated association rules $\boldsymbol{\pi}$, which we call *user association maps*. When a request is generated at a given location, the map probabilistically determines the base station that will serve the user. A *feasible* map $\boldsymbol{\pi}$ must (a) associate to base stations the entire traffic of every location x and (b) ensure through Eq.(3) that the base station loads are limited to < 1 (since $\rho_i > 1$ means that base station i is unstable).

Definition 1 (Feasible user association map)**.**

$$\begin{aligned} \mathcal{F} = \{ & \rho_i \leq 1 - \epsilon,\ \forall i \in \mathcal{B} \\ & \sum_{i \in \mathcal{B}} \pi_i(x) = 1,\ \forall x \in \mathcal{L} \\ & \pi_i(x) \in [0, 1],\ \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \}. \end{aligned} \tag{4}$$

As given in Def.1, $\mathcal{F}$ includes only the most generic and necessary constraints. Our model can be extended to include

application-specific constraints. For example, in the context of UDNs user experience can be improved in crowded locations by further restricting the load of certain base stations, e.g., ensuring $\rho_i < c_i < 1$ for some i; this constraint is handled in this paper. In the context of network slicing, it is useful to define multiple classes of users and design a different map per class such that a user can enjoy a slice-specific QoS level. For clarity of exposition, this extension is left for future work.

Definition 2 (Objective Function)**.** *$\phi(\boldsymbol{\pi})$ is a generic differentiable and seperable convex function.*

Problem 1 (P1: Generic User Association Problem)**.**

$$\underset{\boldsymbol{\pi}\in\mathcal{F}}{\text{minimize}}\ \phi(\boldsymbol{\pi}). \tag{5}$$

The goal of this paper is to provide an algorithm that optimally solves the generic optimization P1 for any choice of $\phi(\boldsymbol{\pi})$. We can later tune the shape of $\phi(\boldsymbol{\pi})$ to drive the system performance according to our objective. We give here some examples: (i) choosing $\phi(\boldsymbol{\pi}) = \sum_i \rho_i$ maximizes the total system throughput, (ii) choosing $\phi(\boldsymbol{\pi}) = \sum_i \rho_i^2$ balances the base station loads, (iii) choosing $\phi(\boldsymbol{\pi}) = \sum_i \rho_i^\alpha,\ \alpha \to \infty$ makes base station loads as equal as possible. In fact, it can be shown that $\phi(\boldsymbol{\pi})$ as given in Definition 2 can be tuned to yield as solution any Pareto-efficient vector $\boldsymbol{\pi}$, and therefore we do not need more generic functions.

Definition 3 (Optimal User Association Map)**.** *A solution of P1 $\boldsymbol{\pi}^\star$ is called the optimal user association map.*

Observe that the optimal solution of P1 strongly depends on knowing the demand $\lambda(x)$ (through Eq.(3)) at a fine spatial granularity x. A number of related works take these as known assuming that they can be estimated from data, cf. [7] and followups. In practice, due to the natural demand fluctuations (especially related to non-stationary phenomena) *there will always be discrepancies between actual and estimated demand, even with the best estimators.*

In the context of UDNs, this poses a great threat to user association, as due to the small number of users per BS, the discrepancies are expected to be larger. To this end, we introduce next our proposed estimators, and then describe how to rigorously treat these unpredicted discrepancies, to avoid violating Service Level Agreements (SLAs).

C. Statistical Methods for Mobile Traffic Prediction

Mobile traffic exhibits strong diurnal patterns, which make it predictable; the interested reader is referred to [10], [11] and [15] for extensive analyses. We use a publicly available dataset collected in the Milano area, analyzed in [11], wherein it has been observed that the daily pattern is stronger when considering each day of the week and each location on the Milano grid separately. Motivated by this we propose the following traffic predictor.

Definition 4 (Traffic Predictor)**.** *Let $X_i^{t,d}(x)$ denote the measured intensity (#of arrivals/h) at location x, hour t, and day d of the week, i weeks before the current. The predicted spatial intensity is based on the data of the last n weeks:*

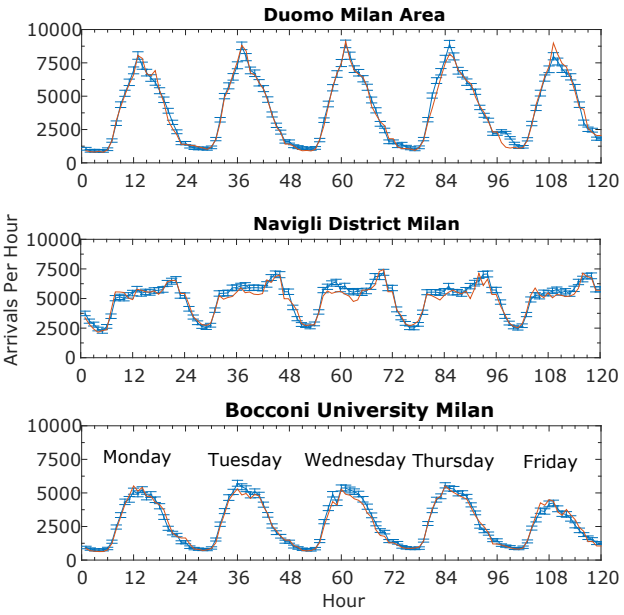


Fig. 2. Traffic Prediction based on Eq.(7) for 3 different areas of Milano from Monday 2/12/2013 to Friday 6/12/2013 and comparison with the actual traffic (a) Duomo Area (b) Navigli District (c) Bocconi University.

$$\bar{\lambda}^{t,d}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^{t,d}(x). \tag{6}$$

Hereinafter, we focus on a single hour/day slot of the week and drop the notation t, d. The actual value of traffic intensity is modeled to be equal to the predicted one $\bar{\lambda}(x)$, plus a zero-mean Gaussian prediction error:

$$\hat{\lambda}(x) = \bar{\lambda}(x) + N_n(x), \tag{7}$$

where $N_n(x) \sim N(0, \sigma_n^2(x))$, and $\sigma_n^2(x)$ is the sample variance, which is given by

$$\sigma_n^2(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i(x) - \bar{\lambda}(x))^2.$$

In Eq.(7) we have implicitly assumed that the observed data of a specific hour of a week are drawn from the same distribution across different weeks, in which case Eq.(7) follows from Eq.(6) and the central limit theorem.

In Fig.2 we evaluate our simple predictor on the Milano dataset [12]. We have trained the predictor for the whole November and used it to predict the first week of December (2-6/12/2013). We see that our model predicts accurately the traffic in most situations for three areas with heterogeneous behavior: 1) the Duomo, with tourist activity 2) the Navigli District, with nightlife activity and 3) the Bocconi University, with working hours activity. More advanced techniques like the ARIMA model [16] can be used for improved predictions, but this is left for future work.

Although the described traffic prediction is fairly accurate, there are spikes in the traffic which are non-stationary and cannot be predicted based on past data, for example see in

Fig.2 at the peak hour on Friday at Duomo (108 hour). If we design the association map disregarding the prediction error, these unpredictable spikes can lead to constraint violations. The next lemma characterizes the behavior of the base station load as a function of the prediction error.

Lemma 1. *Fix (t, d), and let the hourly spatial intensity of traffic $\hat{\lambda}(x)$ be related to the predicted one as explained in Eq.*(7)*. Fix the user association vector $\boldsymbol{\pi}$. The actual load $\hat{\rho}_i$ of base station i is related to the estimated one ρ_i as follows:*

$$\hat{\rho}_i = \rho_i(\boldsymbol{\pi}) + Y_i(\boldsymbol{\pi}), \tag{8}$$

where Y_i is the base station load prediction error, it is zero mean Gaussian random variable with variance:

$$S_i^2(\boldsymbol{\pi}) = \int_{\mathcal{L}} \frac{\pi_i^2(x)}{\mu^2 C_i^2(x)} \sigma_n^2(x) dx. \tag{9}$$

Proof. Analytically, starting from the estimate of $\lambda(x)$:

$$\begin{aligned} \hat{\lambda_n}(x) &= \lambda(x) + N_n(x) \\ \frac{\hat{\lambda_n}(x)}{\mu C_i(x)} &= \frac{\lambda(x)}{\mu C_i(x)} + \frac{N_n(x)}{\mu C_i(x)} \\ \int_{x\in\mathcal{L}} \frac{\hat{\lambda_n}(x)}{\mu C_i(x)} \pi_i(x) dx &= \rho_i + \int_{x\in\mathcal{L}} \frac{N_n(x)}{\mu C_i(x)} \pi_i(x) dx \\ \hat{\rho}_i &= \rho_i + Y_i. \end{aligned} \tag{10}$$

The aggregate noise that is generated from all the locations $x \in \mathcal{L}$ associated with a base station i is described by the random variable: $Y_i = \int_{x\in\mathcal{L}} \frac{N_n(x)}{\mu C_i(x)} \pi_i(x) dx$ and $Y_i \sim N(0, \int_{\mathcal{L}} \frac{\pi_i^2(x)}{\mu^2 C_i^2(x)} \sigma_n^2(x) dx)$, since $N_n(x) \sim N(0, \sigma_n^2(x))$. □

We emphasize that the variance of the error of the predicted load $S_i^2(\boldsymbol{\pi})$ depends on the association rule $\boldsymbol{\pi}$, Eq.(9). Intuitively, the less predicted traffic we associate to a base station, the less confident we are that the actual load will match the predicted, and thus the more conservative we need to be to avoid the violation of its load constraint. This leads us to an interesting observation: *to optimize user association maps under uncertainty we must select the rules $\boldsymbol{\pi}$ considering jointly their impact on the prediction error and the base station load objective.* To capture this tradeoff accurately we introduce the concept of RUAM.

D. Robust User Association Maps

In order to optimize user association maps under uncertainty, we will reformulate P1 using the theory of robust optimization [17],[18]. In terms of objective function, we seek to optimize $\mathbb{E}[\phi(\hat{\rho}_i)]$, where ϕ is the objective function of Definition 2, and the expectation is taken with respect to the Gaussian prediction error. In terms of constraints, we require that the actual BS load does not exceed a tunable parameter c_i with a selected probability ϵ:

$$\text{Prob}(\hat{\rho}_i \geq c_i) \leq \epsilon. \tag{11}$$

Therefore, the *robust* feasibility set $\mathcal{F}^r$ contains association maps $\boldsymbol{\pi}$, such that the predicted base station load $\rho(\boldsymbol{\pi})$ and the prediction error variance $S(\boldsymbol{\pi})$ satisfy certain conditions as explained below.

Problem 2 (P2: Robust User Association Problem)**.**

$$\underset{\boldsymbol{\pi}\in\mathcal{F}^r}{\textit{minimize}}\ \mathbb{E}\left[\phi\left(\rho(\boldsymbol{\pi}) + Y(S(\boldsymbol{\pi}))\right)\right], \tag{12}$$

where $\mathcal{F}^r$:

$$\begin{aligned} \mathcal{F}^r = \{&\textit{Prob}(\rho(\boldsymbol{\pi}) + Y(S(\boldsymbol{\pi})) > c_i) \leq \epsilon \\ &\sum_{i\in\mathcal{B}} \pi_i(x) = 1 \\ &\pi_i(x) \in [0,1]\}. \end{aligned} \tag{13}$$

Definition 5 (Robust User Association Map)**.** *The solution of P2 $\boldsymbol{\pi}^\star$ is called the robust user association map.*

The robust user association map is our novel proposition in this paper. Using past data, we propose to precompute robust maps that will be played out with actual traffic to determine the association at runtime. The maps are easy to use, while at the same time they provide minimum expected cost, and allow us to control the probability of constraint violation. Hence, it is a disciplined and practical approach to optimize the system using data.

Solving P2 is challenging. There is a great number of optimization variables, increasing with the quantization accuracy for area $\mathcal{L}$. Also, in its current form, Eq.(12) is a stochastic program. In the next section we will present an optimal generalized algorithm that overcomes these challenges.

III. Generalized Robust Map Algorithm

We present GRMA: a generalized algorithm for solving the Robust User Association Problem P2 when $\phi(\boldsymbol{\pi})$ is a differentiable and convex objective function. First, we transform the problem into a convex program by replacing the stochastic constraint Eq.(11) with an equivalent convex constraint. Next, we relax the new constraint; the feasibility set of the relaxed problem is a simplex and we can solve it with an efficient projected gradient algorithm. Finally we present the GRMA algorithm, which is based on a dual subgradient method with averaging on the primal sequence $\boldsymbol{\pi}^{(k)}$ and show that it converges to the optimal robust map.

A. Convex Formulation

To make P2 a convex program, we will replace the stochastic constraint $\text{Prob}(\hat{\rho}_i \geq c_i) \leq \epsilon$ in $\mathcal{F}^r$ with an equivalent convex constraint that guarantees protection with ϵ probability.

Lemma 2. *The inequality $\textit{Prob}(\hat{\rho}_i \geq c_i) \leq \epsilon$ is equivalent to $\rho_i + \alpha S_i \leq c_i$, when $\hat{\rho}_i = \rho_i + Y_i$ is taken according to Eq.*(8)*, where Y_i is normally distributed with zero mean and variance and $\alpha = Q^{-1}(\epsilon)$, where $Q(\cdot)$ is the tail probability of the standard normal distribution.*

Proof. Starting by the probabilistic constraint we have:

$$\text{Prob}(\hat{\rho}_i \geq c_i) \leq \epsilon \Leftrightarrow \text{Prob}(Y_i \geq c_i - \rho_i) \leq \epsilon,$$

Y_i is normally distributed with $Y_i \sim N(0, S_i^2)$. If Q is the Q-function (tail probability of the standard normal distribution), we can rewrite the above equation as:

$$Q(\frac{c_i - \rho_i}{S_i}) \leq \epsilon.$$

The inequality is satisfied for all $\frac{c_i-\rho_i}{S_i}$ that:

$$\frac{c_i-\rho_i}{S_i} \geq Q^{-1}(\epsilon) \Leftrightarrow \rho_i + Q^{-1}(\epsilon)S_i \leq c_i.$$

□

From $Q^{-1}(\epsilon)$ it is evident that ϵ is a design parameter that affects the feasibility region $\mathcal{F}^r$. Small values of ϵ protect from violations, but can lead to very inefficient association vectors.

Now, we can define the new set $\mathcal{F}^c$ and prove it is convex.

Definition 6 (Convex Feasibility Set $\mathcal{F}^c$)**.**

$$\begin{aligned}\mathcal{F}^c = \{&\rho_i + \alpha S_i \leq c_i \\ &\sum_{i\in\mathcal{B}} \pi_i(x) = 1 \\ &0 \leq \pi_i(x) \leq 1\}.\end{aligned} \tag{14}$$

Lemma 3. *The constraint $\rho_i + \alpha S_i \leq c_i$ is convex.*

Proof. Consider two vectors $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \mathcal{F}^c$. We first show that $S_i(\theta\boldsymbol{\pi}_1 + (1-\theta)\boldsymbol{\pi}_2) \leq \theta S_i(\boldsymbol{\pi}_1) + (1-\theta)S_i(\boldsymbol{\pi}_2)$, where $\theta \in [0,1]$. Denote $w(x) = \frac{S_n(x)}{\mu C_i(x)}$ We begin by:

$$S_i^2(\theta\boldsymbol{\pi}_1+(1-\theta)\boldsymbol{\pi}_2) = \int w^2(x)(\theta\pi_1(x)+(1-\theta)\pi_2(x))^2 dx$$

$$= S_i^2(\theta\boldsymbol{\pi}_1)+S_i^2((1-\theta)\boldsymbol{\pi}_2)+2\theta(1-\theta)\int w^2(x)\pi_1(x)\pi_2(x)dx$$

and

$$(S_i(\theta\boldsymbol{\pi}_1) + S_i((1-\theta)\boldsymbol{\pi}_2))^2 =$$
$$S_i^2(\theta\boldsymbol{\pi}_1) + S_i^2((1-\theta)\boldsymbol{\pi}_2) + 2S_i(\theta\boldsymbol{\pi}_1)S_i((1-\theta)\boldsymbol{\pi}_2)$$

From the Cauchy-Swartz[1] inequality we have that: $2\theta(1-\theta)\int w^2(x)\pi_1(x)\pi_2(x)dx \leq 2S_i(\theta\boldsymbol{\pi}_1)S_i((1-\theta)\boldsymbol{\pi}_2)$, hence:

$$S_i^2(\theta\boldsymbol{\pi}_1 + (1-\theta)\boldsymbol{\pi}_2) \leq (S_i(\theta\boldsymbol{\pi}_1) + S_i((1-\theta)\boldsymbol{\pi}_2))^2 \Leftrightarrow$$
$$S_i(\theta\boldsymbol{\pi}_1 + (1-\theta)\boldsymbol{\pi}_2) \leq S_i(\theta\boldsymbol{\pi}_1) + S_i((1-\theta)\boldsymbol{\pi}_2) \Leftrightarrow$$
$$S_i(\theta\boldsymbol{\pi}_1 + (1-\theta)\boldsymbol{\pi}_2) \leq \theta S_i(\boldsymbol{\pi}_1) + (1-\theta)S_i(\boldsymbol{\pi}_2).$$

We have proven that S_i is convex, by inspection ρ_i is also convex, and since the sum of positive weighted convex terms is also convex, it follows that the constraint is convex. □

The set $\mathcal{F}^c$ is convex because the constraints in Eq.(14) are convex. Based on Lemmas 2 and 3, and the fact that the expectation of a convex function of a random variable is also convex, we can now recast P2 as a convex program:

Problem 3 (P3: Convex Robust User Association Problem)**.**

$$\underset{\boldsymbol{\pi}\in\mathcal{F}^c}{\textit{minimize}}\ \mathbb{E}\left[\phi\left(\rho(\boldsymbol{\pi}) + Y(S(\boldsymbol{\pi}))\right)\right]. \tag{15}$$

Hence, in the next subsections we focus on resolving the issue of high dimension (great number of variables) optimization with the coupled constraints.

B. Partial Lagrangian Relaxation

The feasiblity set $\mathcal{F}^c$ is convex, but the constraints couple locations and base stations for every association vector and make the implementation of an efficient algorithm challenging. To efficiently solve this, we propose to relax the load constraint. The remaining feasibility set after the relaxation is the simplex $\mathcal{F}' = \{\boldsymbol{\pi} | \sum_{i\in\mathcal{B}} \pi_i(x) = 1, \pi_i(x) \in [0,1]\}$, and there is rich literature on how to apply projected gradient algorithms in simplices, cf. [19],[20]. Our idea here is to keep as many constraints as possible as long as we know how to project infeasible solutions on them, while we relax the rest. Notice that if we would relax all the constraints, we would get an easy unconstrained convex problem, but the coordination of a large number of Lagrangian multipliers would prohibitively delay the solution.

Let us consider the following partial dual maximization, which will be instrumental in solving our problem:

Problem 4 (P4: Partial Dual Robust Problem)**.**

$$\underset{\boldsymbol{\gamma}\geq 0}{\textit{maximize}} \left\{ \underset{\boldsymbol{\pi}\in\mathcal{F}'}{\textit{min}} \{\Phi(\boldsymbol{\pi},\boldsymbol{\gamma}\} \right\}, \tag{16}$$

where the partially relaxed Lagrangian is:

$$\Phi(\boldsymbol{\pi},\boldsymbol{\gamma}) = \mathbb{E}[\phi(\boldsymbol{\pi})] + \sum_{i\in\mathcal{B}} \gamma_i(\rho_i + \alpha S_i - c_i). \tag{17}$$

In Eq.(17) the vector $\boldsymbol{\gamma}$ contains the Lagrangian multipliers. The multipliers penalize association maps which violate the load constraint with extra cost ($\gamma_i \geq 0$), which increases linearly the more overloaded a base station gets. Henceforth we assume that the Slater's Condition holds, which we expect to be the case for all practical purposes in our problem; therefore, the optimal solution of P4 has equal cost with the optimal primal for the P3 (Strong duality [21]).

C. Projected Gradient Descent

In this subsection we design an algorithm to efficiently solve the inner minimization subproblem in Eq.(16). For a given $\gamma^\star$, we have to find the map that minimizes the cost:

$$\underset{\boldsymbol{\pi}\in\mathcal{F}'}{\text{minimize}} \left\{\Phi(\boldsymbol{\pi},\boldsymbol{\gamma}^\star)\right\}. \tag{18}$$

We design the Projected Gradient Descent (PGD) algorithm to solve this problem motivated by the fact that gradient algorithms have been shown in the literature to have independent convergence rate from the dimension (number of variables) of the problem [22, Ch. 3]. Also, the projection onto $\mathcal{F}'$ (simplex) can be solved exactly and efficiently. The algorithm is:

Projected Gradient Descent (PGD) on $\mathcal{F}'$

Initialize: $\boldsymbol{\pi}^{(0)}$ (can be infeasible), $\boldsymbol{\gamma}^\star$.
Iterate: over n, until convergence

$$\boldsymbol{y}^{(n+1)} = \boldsymbol{\pi}^{(n)} - s^{(n)}\nabla_{\boldsymbol{\pi}}\Phi(\boldsymbol{\pi}^{(n)},\boldsymbol{\gamma}^\star) \tag{19}$$
$$\boldsymbol{\pi}^{(n+1)} = \Pi_{splx}[\boldsymbol{y}^{(n+1)}] \tag{20}$$

Where Π_{splx} is the projection on $\mathcal{F}'$:

Sort $\boldsymbol{y}^{(n+1)}$ in descending order ($y_1 \geq y_2 \geq \ldots \geq y_{|\mathcal{B}|}$)
Select $m = \underset{j\in\mathcal{B}}{\text{argmax}}\{j \mid y_j + \frac{1}{j}(1 - \sum_{i=1}^{j} y_i) > 0\}$
$\pi_i^{(n+1)} = \left[y_i + \frac{1}{m}(1 - \sum_{i=1}^{m} y_i)\right]^+,\ i = 1,\ldots,|\mathcal{B}|$

Eq.(19) implements the gradient update of the user association $\boldsymbol{\pi}^{(n)}$ one step along the direction of the gradient with fixed step size $s^{(n)}$. Eq.(20) is the orthogonal projection of $\boldsymbol{y}^{(n+1)}$

[1]Define $f(x) = w(x)\pi_1(x)$ and $g(x) = w(x)\pi_2(x)$, then we have that $|\int f(x)g(x)dx|^2 \leq \int |f(x)|^2 dx \int |g(x)|^2 dx$

onto the set $\mathcal{F}'$, which is a simplex. Π_{splx}, as described here, is shown in [20] to give an exact solution to the projection in $\mathcal{O}(|\mathcal{B}|\log|\mathcal{B}|)$.

Proposition 1 (Convergence Rate of PGD)**.** *Let $\boldsymbol{\pi}^{(n)}$ be the projected output of PGD algorithm at iteration n, and $\boldsymbol{\pi}^\star$ be an optimal solution of* (18)*, it is shown in [22, Ch. 3.2]:*

$$||\boldsymbol{\pi}^{(n)} - \boldsymbol{\pi}^\star|| = \mathcal{O}(1/n).$$

D. Dual Subgradient Method

We return to the task of solving P4. The objective of this problem $D(\boldsymbol{\gamma}) = \min_{\boldsymbol{\pi}\in\mathcal{F}'}\{\Phi(\boldsymbol{\pi},\boldsymbol{\gamma})\}$ is not γ-differentiable everywhere, hence we will resort to a subgradient method [23],[24] for updating the value of the multipliers.

Proposition 2 (Subgradient Vector)**.** *The vector:*

$$\boldsymbol{g}^{(k)} = \rho(\boldsymbol{\pi}^{(k)}) + \alpha S(\boldsymbol{\pi}^{(k)}) - \boldsymbol{c},$$

where

$$\boldsymbol{\pi}^{(k)} \in \underset{\boldsymbol{\pi}\in\mathcal{F}'}{argmin}\left\{\mathbb{E}[\phi(\boldsymbol{\pi})] + \sum_{i\in\mathcal{B}}\gamma_i^{(k)}(\rho_i + \alpha S_i - c_i)\right\}, \quad (21)$$

satisfies

$$||D(\boldsymbol{\gamma}^{(k+1)}) - D(\boldsymbol{\gamma}^{(k)})|| \le \boldsymbol{g}^{(k)}||\boldsymbol{\gamma}^{(k+1)} - \boldsymbol{\gamma}^{(k)}||$$

as shown in[23, Ch. 6.1], hence is a subgradient of $D(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma}^{(k)}$.

We now show that the norm of the subgradients is bounded; a necessary property for the convergence of the method.

Lemma 4 (Bounds on the Subgradient)**.** *The subgradient sequence $\{\boldsymbol{g}^{(k)}\}$ is bounded:*

$$||\boldsymbol{g}^{(k)}|| < L, \quad (22)$$

where

$$L = \sqrt{\sum_{i\in\mathcal{B}}\left(\int_{x\in\mathcal{L}}\frac{\lambda(x)}{\mu C_i(x)}dx + \alpha\int_{x\in\mathcal{L}}\frac{\sigma_n^2(x)}{\mu^2 C_i(x)^2}dx\right)^2}. \quad (23)$$

Proof. The set $\mathcal{F}'$ is a simplex (compact). The constraints $\boldsymbol{g}_i,\ i\in\mathcal{B}$ are convex over $\mathbb{R}^n$, hence they are continuous over $\mathbb{R}^n$. The norm of the subgradients is upper bounded by $\max_{\boldsymbol{\pi}\in\mathcal{F}'}||g(\boldsymbol{\pi})||$. This is smaller than assigning all the locations $x\in\mathcal{L}$ to all base stations, hence the norm of the subgradients is bounded by the easy to calculate Eq.(23). □

The subgradient method updates the multipliers $\boldsymbol{\gamma}^{(k)}$ by making a step along the direction of the subgradient vector:

$$\boldsymbol{\gamma}^{(k+1)} = [\boldsymbol{\gamma}^{(k)} + s^{(k)}\boldsymbol{g}^{(k)}]^+. \quad (24)$$

For dual problems with a unique solution, the above algorithm converges to the unique optimal dual vector $\boldsymbol{\gamma}^\star$, and with this we can calculate the optimal robust map $\boldsymbol{\pi}^\star$ by a single run of PGD algorithm. However, P4 is not strictly convex, and therefore its dual may have multiple solutions. The subgradient method may converge to a solution $(\boldsymbol{\pi},\boldsymbol{\gamma})$ which does not satisfy complementary slackness and hence $\boldsymbol{\pi}$ is not feasible in P3 (it will violate the load constraint). To alleviate this issue we will use the technique of averaging: the idea is to output as a solution the average of the primal iterates $\boldsymbol{\pi}^{(k)}$ (feasible or not). We will show that the sequence of averages $\bar{\boldsymbol{\pi}}^{(k)}$ converges to the optimal solution of P3.

Generalized Robust Map Algorithm (GRMA)

Initialize: $\boldsymbol{\pi}^{(0)}$ (e.g. *max-SINR*, can be infeasible), $\boldsymbol{\gamma}^{(0)}$.
Iterate: over k, until convergence
$\boldsymbol{\gamma}^{(k+1)} = [\boldsymbol{\gamma}^{(k)} + s^{(k)}\boldsymbol{g}^{(k)}]^+$
$\boldsymbol{\pi}^{(k+1)} \leftarrow \text{PGD}(\boldsymbol{\gamma}^{(k+1)})$
Keep the running average of the $\boldsymbol{\pi}^{(k)}$ (Eq.(21)*):*
$\bar{\boldsymbol{\pi}}^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\pi}^{(i)}$

Theorem 1 (Convergence to Primal Optimal)**.** *The average of the primal iterates $\bar{\boldsymbol{\pi}}^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\pi}^{(i)}$, where*

$$\boldsymbol{\pi}^{(i)} \in \underset{\boldsymbol{\pi}\in\mathcal{F}'}{argmin}\left\{\mathbb{E}[\phi(\boldsymbol{\pi})] + \sum_{j\in\mathcal{B}}\gamma_j^{(i)}(\rho_j + \alpha S_j - c_j)\right\}, \quad (25)$$

asymptotically converges to (or approximates) the optimal robust association map $\boldsymbol{\pi}^\star$, i.e.:

$$\lim_{k\to\infty}||g(\bar{\boldsymbol{\pi}}^{(k)})^+|| \to 0 \text{ and } \lim_{k\to\infty}\phi(\bar{\boldsymbol{\pi}}^{(k)}) = \phi(\boldsymbol{\pi}^\star).$$

Proof. We use a constant step size, hence $s^{(k)} = s$. We also denote $\boldsymbol{\gamma}^\star$ as the optimal multipliers and

$$d^\star = \max_{\boldsymbol{\gamma}\ge 0}\left\{\min_{\boldsymbol{\pi}\in\mathcal{F}'}\{\Phi(\boldsymbol{\pi},\boldsymbol{\gamma}\}\right\}$$

is the optimal value of the dual problem. First we prove that the load constraint violation for the vector $\bar{\boldsymbol{\pi}}^{(k)}$ is upper bounded as follows:

$$||g(\bar{\boldsymbol{\pi}}^{(k)})^+|| \le \frac{||\boldsymbol{\gamma}^{(k)}||}{ks}. \quad (26)$$

By updating the dual as described in Eq.(24), we have:

$$sg(\boldsymbol{\pi}^{(k)}) \le \boldsymbol{\gamma}^{(k+1)} - \boldsymbol{\gamma}^{(k)},\ \forall k \ge 0.$$

Summing telescopically for $i = 0, 1, \ldots, k-1$ we get:

$$\sum_{i=0}^{k-1} sg(\boldsymbol{\pi}^{(i)}) \le \boldsymbol{\gamma}^{(k)} - \boldsymbol{\gamma}^{(0)} \le \boldsymbol{\gamma}^{(k)},\ \forall k \ge 1. \quad (27)$$

Also, since $g(\bar{\boldsymbol{\pi}}^{(\boldsymbol{k})})$ is convex, we have that:

$$g(\bar{\boldsymbol{\pi}}^{(\boldsymbol{k})}) = g(\frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\pi}^i) \le \frac{1}{k}\sum_{i=0}^{k-1}g(\boldsymbol{\pi}^{(i)})$$
$$= \frac{1}{ks}\sum_{i=0}^{k-1}sg(\boldsymbol{\pi}^{(i)}) \overset{Eq.(27)}{\le} \frac{\boldsymbol{\gamma}^{(k)}}{ks}.$$

Taking norms for the active constraints ($g(\bar{\boldsymbol{\pi}}^{(k)}) \ge 0$) gives Eq.(26). Since the Lagrangian multipliers are bounded [24, Lem.3], the first result follows. Next, we will prove that the objective function for the vector $\bar{\boldsymbol{\pi}}^{(k)}$ is upper bounded by:

$$\phi(\bar{\boldsymbol{\pi}}^{(k)}) \le d^\star + \frac{||\boldsymbol{\gamma}^{(0)}||^2}{2ks} + \frac{s}{2k}\sum_{i=0}^{k-1}||g(\boldsymbol{\pi}^{(i)})||^2. \quad (28)$$

(a) (b) (c) (d) (e)

Fig. 3. Association Maps (a) Voronoi Cells (b) GRMA convergence for max rate (c) robust maximum rate map ($\epsilon = 0.05$) (d) GRMA convergence for proportional fairness (e) robust proportional fair map ($\epsilon = 0.05$).

From Eq.(24):

$$||\boldsymbol{\gamma}^{(i+1)}||^2 \leq ||\boldsymbol{\gamma}^{(i)}||^2 + s^2||g(\boldsymbol{\pi}^{(i)})||^2 + 2s\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)}) \Leftrightarrow$$

$$-\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)}) \leq \frac{||\boldsymbol{\gamma}^{(i)}||^2 - ||\boldsymbol{\gamma}^{(i+1)}||^2 + s^2||g(\boldsymbol{\pi}^{(i)})||^2}{2s}.$$

By taking the telescoping sum we have:

$$-\frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)}) \leq \frac{||\boldsymbol{\gamma}^{(0)}||^2 - ||\boldsymbol{\gamma}^{(k)}||^2}{2ks} + \frac{s}{2k}\sum_{i=0}^{k-1}||g(\boldsymbol{\pi}^{(i)})||^2$$

$$\leq \frac{||\boldsymbol{\gamma}^{(0)}||^2}{2ks} + \frac{s}{2k}\sum_{i=0}^{k-1}||g(\boldsymbol{\pi}^{(i)})||^2. \tag{29}$$

As before, since $\phi(\boldsymbol{\pi})$ is convex:

$$\phi(\bar{\boldsymbol{\pi}}^{(k)}) = \phi(\frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\pi}^{(i)}) \leq \frac{1}{k}\sum_{i=0}^{k-1}\phi(\boldsymbol{\pi}^{(i)})$$

$$\overset{Eq.(30)}{\leq} \frac{1}{k}\sum_{i=0}^{k-1}D(\boldsymbol{\gamma}^{(i)}) - \frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)})$$

$$\leq d^\star - \frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)})$$

$$\overset{Eq.(29)}{\leq} d^\star + \frac{||\boldsymbol{\gamma}^{(0)}||^2}{2ks} + \frac{s}{2k}\sum_{i=0}^{k-1}||g(\boldsymbol{\pi}^{(i)})||^2.$$

The second inequality is true because $\boldsymbol{\pi}^{(i)}$ is a minimizer of the Lagrangian Eq.(25) and:

$$\phi(\bar{\boldsymbol{\pi}}^{(k)}) \leq \frac{1}{k}\sum_{i=0}^{k-1}\Phi(\boldsymbol{\pi}^{(i)}, \boldsymbol{\gamma}^{(i)}) - \frac{1}{k}\sum_{i=0}^{k-1}\boldsymbol{\gamma}^{(i)}g(\boldsymbol{\pi}^{(i)}). \tag{30}$$

By taking $k \to \infty$ on Eq.(26) shows that $\bar{\boldsymbol{\pi}}^{(k)}$ is feasible and on Eq.(28) shows that $\phi(\bar{\boldsymbol{\pi}}^{(k)}) \to d^\star$. The result follows. □

E. Example Applications of GRMA

First, we consider the maximum expected rate objective, where the optimal map will associate every location x to the base stations that provide the highest physical rate $C_i(x)$. This, according to Eq.(3) is identical to minimizing the sum of loads. Hence, taking expectation of the cost of the predicted $\hat{\rho}$:

$$\mathbb{E}\left[\sum_{i\in\mathcal{B}}\hat{\rho}_i\right] = \sum_{i\in\mathcal{B}}\mathbb{E}[\rho_i + Y_i] = \sum_{i\in\mathcal{B}}\rho_i.$$

Robust Map 1 (RM1: Maximum Expected Rate Map)**.** *The optimal maximum expected rate map $\pi^\star$ is the solution of:*

$$\underset{\boldsymbol{\pi}\in\mathcal{F}^c}{minimize}\{\sum_{i\in\mathcal{B}}\rho_i\}. \tag{31}$$

TABLE I
SIMULATION PARAMETERS [27]

Parameter	Variable	Value
Transmission Power Macro BS	P_M	43 dbm
Transmission Power Micro BS	P_m	33 dbm
System Bandwidth	W	10 MHz
Noise Density	No	-174dbm/Hz
Path Loss Exponent	P_{lo}	3

Next we consider the *penalty proportional fairness* associated to the objective $\phi(\boldsymbol{\pi}) = \sum_i \frac{\rho_i^2}{2}$ [25], [26]. This condition leads to a load balancing trade-off, where base stations in high traffic areas are allowed to be more loaded in the benefit of higher total throughput. Taking expectation:

$$E_Y\left[\sum_{i\in\mathcal{B}}\frac{(\rho_i + Y_i)^2}{2}\right] = \sum_{i\in\mathcal{B}}\frac{\mathbb{E}[\rho_i^2 + S_i^2 + 2\rho_i S_i]}{2}$$

$$= \sum_{i\in\mathcal{B}}\frac{\rho_i^2 + S_i^2}{2}.$$

Robust Map 2 (RM2: Proportional Fair Map)**.** *The optimal proportional fair map $\pi^\star$ is given by:*

$$\underset{\boldsymbol{\pi}\in\mathcal{F}^c}{minimize}\left\{\sum_{i\in\mathcal{B}}\frac{\rho_i^2 + S_i^2}{2}\right\}. \tag{32}$$

Figure 3 shows the progress in iterations towards convergence and feasibility of GRMA and an illustration of the optimal robust map produced when applied on a network setup of heterogeneous base stations in an area with highly variable and dense traffic. In Fig.3b and 3d, we can see that for both objectives GRMA, after 200 iterations, produces feasible solutions with almost optimal cost. In the maximum rate map 3c, we can see the similarities with Voronoi cells 3a, at locations where the expected traffic is low, while at heavy traffic locations we have curved boundaries, enforced by the load protection constraint. In proportional fair map 3e, the cells are very different from the other two cases.

IV. NUMERICAL EVALUATION

A. Simulation Setup

Here, we will compare the proposed robust user association maps to an adaptive version of a popular user association algorithm from the literature [7], with the same optimization objective. At every network update (10 minutes), the adaptive algorithm calculates the average load experienced on the

previous slot and settles to a new association vector, while on the other hand we apply our precalculated map for every hour. The accrued cost is the value of the objective function, based on the actual input at that time slot and on the user association policies currently active. The average delay is the average response time for a flow in the network. Finally, we count violations as the percentage of time in which the system has an or some overloaded base stations; this happens when either an SLA with a load threshold c_i is violated or when some of the base stations are overcumbered by traffic ($\rho_i > 1$).

We will experiment on the Milano dataset [12], which provides spatially aggregated data about the telecommunication activity. The data are grouped on a regular grid overlaying the territory of Milano with 100×100 squares. Consequently, the grid designates the area $\mathcal{L}$ and every square is a location $x \in \mathcal{L}$ to be associated with base stations. For every square of this grid the data set contains the aggregate per ten minutes telecommunication events in the period of 01/11/13-01/01/14. In this work we consider weekdays (Monday to Friday) which are non-holidays, since then the volume of traffic is increased. We want to emphasize that our framework is especially effective for holidays, and other rare but predictable occasions, like a football match or a concert, for which network operators can reserve a special map, tuned to an exceptional increase/decrease in predicted traffic.

We choose an architected setup of 40 base stations with fixed positions, spread over the area, with higher density on the area that is the city center. We specifically design this subset of base stations, to accurately simulate a simplified environment of a UDN, bringing in the front all the aspects of the user association problem. In the simulations scenarios we will consider two alternative base station setups. One, in which all the traffic is served by the small cells, as envisioned for future 5G UDNs, cf. [1], and one is with the two tier structure, which is dominant in current networks. The LTE parameters used in the simulation are given in table I.

B. Robust Maps vs Adaptive Algorithm

In the first experiment we focus on the choice of ϵ and the effect it has to the performance of a robust map. In theory the choice of smaller ϵ shrinks the feasibility space, allowing only maps that provide an ϵ probability protection guarantee (Eq.(11)). This should correlate with the violation metric in the results and also we expect slightly increased cost due to eliminating cheaper but more risky configurations. This behavior is well observed in the results.

In tables II, III, we present performance results for different values of ϵ during rush hours and in general. We observe that the robust maps typically incur a small increase of cost ($< 10\%$ in average, and $< 18\%$ during peak hours) with respect to the adaptive algorithm. On the other hand, the robust approach provides extraordinary guarantees against traffic fluctuations. In particular, we observe a significantly better average delay ($\approx 30\%$ better) and much less violations (0 instead of 25% of the adaptive algorithm). We also observe that selecting a more relaxed $\epsilon = 5\%, 10\%$, reduces the cost, but deteriorates the performance with respect to average delay and violations of the max load in the duration of the simulation.

TABLE II
AGGREGATE RESULTS 1ST WEEK OF DECEMBER MICRO SETUP

ϵ	Average Cost	Average Delay (s)	Violations (%)
Adaptive	6.001	1.700	13.1
10%	6.462	1.353	2.3
5%	6.485	1.319	1.6
0.1%	6.596	1.267	0

TABLE III
PEAK TRAFFIC 1ST WEEK OF DECEMBER MICRO SETUP

ϵ	Average Cost	Average Delay (s)	Violations (%)
Adaptive	9.671	2.858	24.6
10%	10.911	2.251	7.7
5%	11.000	2.140	3.1
0.1%	11.415	2.030	0

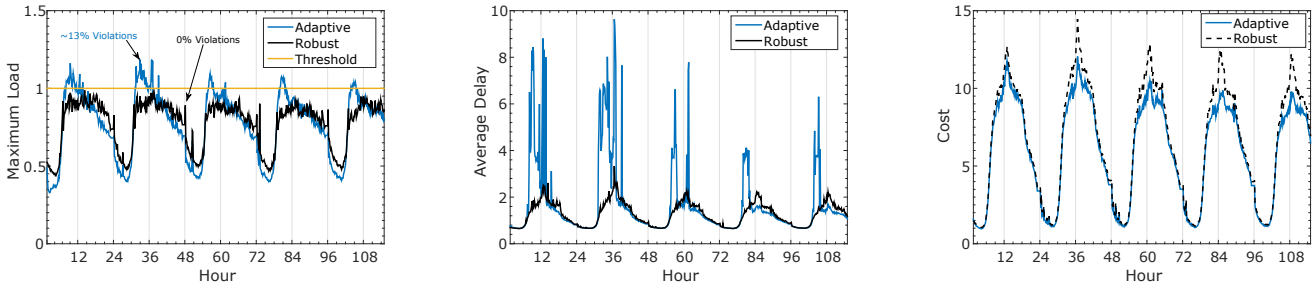
Moreover, in Fig.4 we depict the time evolution of the system under the two considered approaches (the robust map and the adaptive algorithm). Comparing the two approaches in this scenario, we see that the robust maps yield 0% violations vs 13% of the adaptive algorithm, this improvement leads to a much better delay performance. Notice, that the spikes in average delay in Fig.4(b) correspond to load constraint violations in Fig.4(a). Additionally, the improvements in the performance guarantees come at a very small cost increase Fig.4(c). Remarkably our scheme incurs no extra cost when the traffic is low, a benefit that arises from using different maps per hour and exploiting the past data for prediction.

Last, Fig.5 presents a scenario in which we enforce SLAs, in the form of a cap on the base station loads ($\rho_i \leq 0.9$). In practice, SLA violations are extremely important and must be avoided at all costs. The results of Fig.5 show how our robust maps ($\epsilon = 5\%$) protect the SLA from violations, resulting in only 4% violations instead of 15% of the adaptive algorithm; this would be further improved by a more conservative ϵ.

In both time evolution figures we can see that the adaptive algorithms have a natural way of adapting to fluctuations, however this takes a lot of time and in the meantime the system tends to exhibit unstable behavior. Finally, the robust maps pay an increased optimization cost (we have already argued that is of lesser significance than SLA failures) for being conservative against these failures. This is more evident during peak traffic hours, where the more conservative handling of the robust maps infers greater cost, but also improved delay performance and protection against the fluctuating traffic, for example see Fig.4b and Fig.4a arround hour 36.

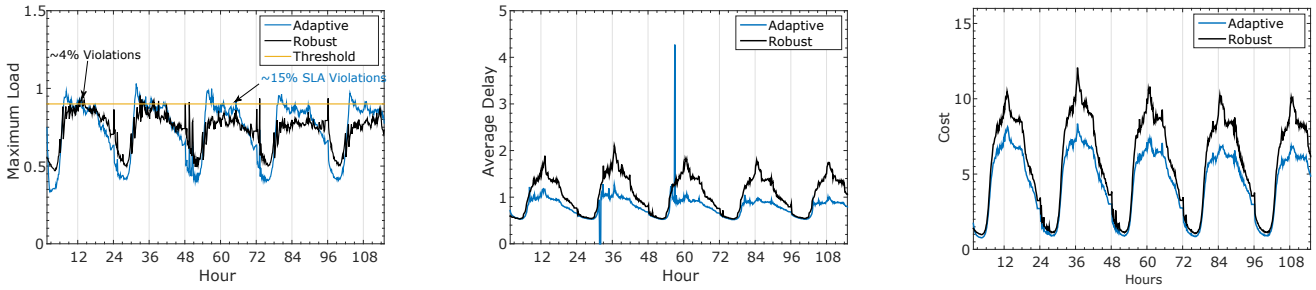
V. CONCLUSION

The problem of user association in dense networks becomes challenging due to frequent unexpected traffic fluctuations. We showed that past traffic data can be exploited towards precalculating association maps, which are designed to be robust and can be tuned to protect the base stations from overload. Accordingly, we proposed a theoretical framework for efficiently computing the optimal robust map, parametrized



(a) (b) (c)

Fig. 4. Micro Base Station Setup, strong SLA protection $\epsilon = 0.001$, 2-6 of December (a) Violation (b) Average System Delay (c) Cost



(a) (b) (c)

Fig. 5. 2-Tier Base Station Setup, light SLA protection $\epsilon = 0.05$, 2-6 of December (a) Violation (b) Average System Delay (c) Cost

to a large class of utility functions that allow the system designer to tune the base station load. Finally, we evaluated our approach in Milano dataset, and found that our methodology is very effective at protecting UDNs from unexpected spikes, allowing the offering of premium wireless service.

REFERENCES

[1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[2] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[3] Nokia, "Ultra Dense Network UDN," 2016.

[4] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User Association in 5G Networks: A Survey and an Outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, 2016.

[5] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[6] T. Bu, L. Li, and R. Ramjee, "Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks," in *Proceedings IEEE INFOCOM*, April 2006, pp. 1–12.

[7] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha-Optimal User Association and Cell Load Balancing in Wireless Networks," *IEEE/ACM Transactions on Networking*, Feb 2012.

[8] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal Downlink and Uplink User Association in Backhaul-limited HetNets," in *IEEE INFOCOM*, 2016.

[9] D. Bertsimas, V. Gupta, and N. Kallus, "Data-driven Robust Optimization," *Mathematical Programming*, Feb 2017.

[10] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *IEEE/ACM Transactions on Networking*, vol. PP, no. 99, pp. 1–15, 2016.

[11] G. Barlacchi et al., "A Multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino," 2015. [Online]. Available: http://dx.doi.org/10.1038/sdata.2015.55

[12] Telecom Italia, "Telecommunications - SMS, Call, Internet - MI," 2015. [Online]. Available: http://dx.doi.org/10.7910/DVN/EGZHFV

[13] K. Shen and W. Yu, "Distributed Pricing-Based User Association for Downlink Heterogeneous Cellular Networks," *CoRR*, 2014.

[14] T. Bonald and A. Proutière, "Wireless Downlink Data Channels: User Performance and Cell Dimensioning," in *MobiCom 2003*. ACM.

[15] D. Naboulsi, M. Fiore, and R. Stanica, "Human Mobility Flows in the City of Abidjan," 2013. [Online]. Available: https://hal.inria.fr/hal-00908277

[16] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer, 2016.

[17] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton University Press, 2009.

[18] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and Applications of Robust Optimization," *SIAM review*, vol. 53, no. 3, pp. 464–501, 2011.

[19] Y. Chen and X. Ye, "Projection Onto A Simplex," 2011.

[20] W. Wang and M. Á. Carreira-Perpiñán, "Projection onto the Probability Simplex: An Efficient Algorithm with a Simple Proof, and an Application," 2013. [Online]. Available: http://arxiv.org/abs/1309.1541

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[22] S. Bubeck *et al.*, "Convex Optimization: Algorithms and Complexity," *Foundations and Trends in Machine Learning*, 2015.

[23] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.

[24] A. Nedić and A. Ozdaglar, "Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.

[25] C. p. Li and M. J. Neely, "Delay and Rate-Optimal Control in a Multi-Class Priority Queue with Adjustable Service Rates," in *Proceedings IEEE INFOCOM*, March 2012.

[26] C. p. Li, G. S. Paschos, L. Tassiulas, and E. Modiano, "Dynamic Overload Balancing in Server Farms," in *IFIP Networking Conference*, June 2014, pp. 1–9.

[27] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An Analytical Framework for Optimal Downlink-Uplink User Association in HetNets with Traffic Differentiation," in *IEEE GLOBECOM*, 2015.