# An Analytical Framework for Optimal Downlink-Uplink User Association in HetNets with Traffic Differentiation

Nikolaos Sapountzis[1], Thrasyvoulos Spyropoulos[1], Navid Nikaein[1], and Umer Salim[2]

[1]Mobile Communications Department, EURECOM, 06410, Biot, France, firstname.lastname@eurecom.fr
[2]Intel Mobile Communications, Sophia Antipolis, 06560, France, umer.salim@intel.com

***Abstract*—The widespread adoption of tablets and smartphones, and an abundance of data-hungry mobile applications, are overwhelming wireless networks with increased demand and introduce considerable traffic diversity. Operators struggling to continuously add capacity and upgrade their architecture have resorted instead to building denser deployments to improve spectral efficiency. By increasing the number of cells a user can associate with, (i) user quality of service (QoS) can be improved, and (ii) traffic can be offloaded from congested base stations, to achieve better load balancing. However, these two goals are not always aligned. To this end, we develop an analytical framework for optimal user association in future HetNets that investigates the potential tradeoffs between user- and network-related performance, in a more realistic setup encompassing additional key features: (i) different types of user flows, and (ii) uplink and downlink performance. We believe this better reflects the diversity of the services offered to users and their impact on system performance. We evaluate our proposed framework through extensive simulations, and provide some qualitative and quantitative insights on the related tradeoffs.**

## I. Introduction

Lately, heterogeneous networks (HetNets) have been widely discussed in the LTE-A (Long Term Evolution - Advanced) [1]. In a HetNet, *small cells* (SC) are deployed along with macrocells to improve spatial reuse, and provide additional capacity in areas with dense usage (i.e., "hotspots"), such as train stations, airports or malls. The higher the deployment density, the better the chance that a user equipment (UE) can be associated with a nearby base station (BS) with high signal strength, and the more the options to balance the load.

However, denser deployments experience high spatio-temporal load variations, and require more sophisticated load-balancing and user association algorithms [2]. Addressing such issues becomes even more challenging when one considers HetNets, i.e. networks consisted of BSs with different transmit powers. There are two key concerns when assigning a UE to a BS: (i) choosing the BS that maximizes the QoS for this user (e.g. the *physical date rate*); (ii) ensuring that the load across BSs is balanced, to avoid congestion. We will refer to the former as the *user-perspective* and the latter as the *network-perspective*[1]. These two goals are often conflicting. Standard SINR-based association might lead a UE to choose a high-power or nearby BS, to maximize it's rate, but this BS might already be congested on the radio or backhaul link [3]. A recently proposed framework [4] makes an important first step towards investigating this relationship between user- and network-related performance, when performing user association. A parameterized objective is used which jointly captures both metrics, with a parameter $\alpha$ controlling which of the two objectives carries more importance. An optimal association rule is then derived for this objective.

Nevertheless, the above framework [4], as well as a number of follow up works in this context, are relatively simplified, not taking into account key features of future networks. Firstly, most existing studies consider homogeneous traffic profiles. For example, [4], [5], [6] assume that all flows generated by a UE are "best-effort" (or "elastic"). Modern and future networks will have to deal with high traffic differentiation, with certain flows being able to require specific, *dedicated* resources [1], [7], [8]. Such dedicated flows do not "share" BS resources like best-effort ones, are sensitive to additional QoS metrics, and affect cell load differently.

Additionally, the majority of related studies consider downlink (DL) traffic only [4], [5], [9]. A user-association criterion that takes into account only the uplink (UL) or only the DL is not sufficient according to [10]. More precisely, the asymmetric transmit powers between the UEs and different BSs differentiate the DL and UL physical data rates significantly. To this end, associating a UE with the BS that offers the highest DL SINR, may lead to subpar UL performance or require high UE transmission power. What is more, the traffic load on the DL and UL may vary significantly, due to the asymmetric traffic applications [11]. For instance, when a user is browsing he consumes resources mostly from the downlink, when uploading a video from the uplink, or when playing an online interactive video game from both downlink and uplink. Summarizing, a proper user-association scheme becomes even more complex if one considers the user and network performance in the DL and the UL *jointly*.

To this end, we revisit the problem of user association in a more complex setup. We use the basic methodology proposed in [4] as our starting point, and extend the framework considerably, to include these key additional dimensions, namely traffic heterogeneity, and differentiation in UL and DL traffic. Specifically, our contributions can be summarized as follows:

This work was supported by the European Research Council under the European Community Seventh Framework Programme (FP7/2012- 2015) under the ICT theme of DG-CONNECT $n^{o}$ 317941 (iJOIN).

[1]*User* performance can also be affected by congestion, when a BS is overloaded. However, we make this simple distinction to facilitate our discussion.

1) We introduce dedicated flows into the framework, along with a different scheduling discipline and QoS metrics; an optimal rule can still be derived when jointly considering user- and network-related performance for both types of flows.

2) We take into account the differentiation between DL and UL traffic and prove that an optimal association rule can also be derived that jointly considers DL and UL performance.

3) We show that our framework also applies when UL and DL traffic of the same UE can be "split" to different BSs [12], as a disruptive architectural design for future 5G networks [13].

4) We include all the above features into the cost function, and prove an optimal rule for the complete setting. Interestingly, the optimal rule when considering multiple objectives resembles a (weighted) harmonic mean of the individual association rules.

5) We further investigate the complex tradeoffs involved *quantitatively* to provide some initial insights and guidelines about user-association policies in future HetNets, and sketch a potential implementation of our algorithm using a Software Defined Network (SDN) architecture.

The remainder of the paper is organized as follows: Section II outlines the considered scheduling disciplines and our system model. The proposed framework for the optimal user-association is described in Section III, and a flexible SDN implementation architecture in Section IV. Section V presents some simulation results, and Section VI concludes the paper.

## II. System Model

Throughout this paper we assume a region $\mathcal{L} \subset \mathbb{R}^2$ served by a set of BSs $\mathcal{B}$, that are either macro BSs (eNBs) or small cells (SCs). We use $i \in \mathcal{B}$ to index a typical $i$-th BS. We let $x \in \mathcal{L}$ denote a location where a User Equipment (UE) is located and a flow might initiate from. Moreover, we sketch the traffic arrival and service models, with respect to (wrt) the additional dimensions discussed earlier. In Table I, we summarize some useful notation we use throughout the paper.

### A. Traffic Arrival Model

To model the spatial traffic variation, we assume that flow arrivals follow an inhomogeneous Poisson Point Process (PPP) with total arrival rate $\lambda(x)$ per unit area. Each new flow is [1]:

- a **downlink (DL)** flow with probability $z_{DL}$, with direction from the BS to the UE, or
- an **uplink (UL)** flow with probability $z_{UL} = 1 - z_{DL}$, with direction from the UE to the BS, independently.

Each DL (or UL) flow is also, independently [1], [7], [8]:

- a **dedicated** flow with probability $z_d = 1 - z_b$; *dedicated* bearers are allocated for Guaranteed Bit Rate (GBR) type of traffic to meet the required bit rate or latency constraints. These are differentiated by their QoS class of identifier (QCI) ranging from 1 to 4 [1],
- a **best-effort** flow, with probability $z_b$, related to non-GBR traffic, and QCI from 5 to 9 [1].

The parameters $z_{DL}$ and $z_b$ depend on the traffic mix, and we assume them to be input parameters. Using the Poisson splitting argument [14], it follows that there are 4 independent, Poisson flow arrival processes with respective rates

$$\lambda_1(x) = z_{DL} \cdot z_b \cdot \lambda(x), \quad \lambda_2(x) = z_{DL} \cdot z_d \cdot \lambda(x) \tag{1}$$

$$\lambda_3(x) = z_{UL} \cdot z_b \cdot \lambda(x), \quad \lambda_4(x) = z_{UL} \cdot z_d \cdot \lambda(x). \tag{2}$$

Throughout the paper, we use indices 1,2,3,4 to refer to the following flow types: DL best effort (1), DL dedicated (2), UL best effort (3), and UL dedicated (4), respectively. Finally, Fig. 1 depicts their corresponding scheduling disciplines that we elaborate on, in the remainder of this section.

### B. Service model for best-effort flows

Best-effort flows are statistically multiplexed and have to compete for resources. A lot of effort has been devoted to the study of the performance and scheduling algorithms for such "elastic" types of traffic [14] [15].

We start with a simple scenario of a single DL (resp. UL) best-effort flow alone in the cell, requested by a UE at location $x$. Its received signal to interference plus noise ratio (SINR) is

$$\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j \neq i} G_j(x)P_j + N_0}, \tag{3}$$

where $N_0$ is the noise power, $P_i$ the transmission power of BS $i$, and $G_i(x)$ represents the path loss and shadowing effects between the i-th BS and the UE located at $x$ (it may also encompass antenna and coding gains etc). We assume that effects of fast fading are filtered out [4], [5].

We assume further that the available bandwidth of the $i$-th BS in the DL is $w_i$, and it is allocated between best effort and dedicated flows, as follows [8]: $\zeta_i \cdot w_i$ is the bandwidth allocated for best-effort flows and $(1 - \zeta_i) \cdot w_i$ for dedicated flows, respectively $(0 \leq \zeta_i \leq 1)$. We can use Shannon's formula to derive the *physical data rate* for DL best-effort flows at $x$:

$$c_{\text{i,1}}(x) = \zeta_i \cdot w_i \log_2(1 + \text{SINR}_i(x)). \tag{4}$$

We similarly assume that the available UL bandwidth at BS $i$ is $W_i$ and further split between UL best effort and dedicated flows wrt another parameter. Hence, the physical rate for UL best-effort traffic is[2] $c_{\text{i,3}}(x) = Z_i \cdot W_i \log_2(1 + \text{SINR}_i(x))$ at $x$. Regarding the single DL flow, or single user, case, $c_{\text{i,1}}(x)$ is the effective service rate (resp. $c_{\text{i,3}}(x)$ for UL). The *user-perspective*, wrt the DL traffic, corresponds to attaching to the BS $i$ that maximizes the rate $c_{i,1}(x)$ (or $c_{i,3}(x)$ in the UL).

However, if there are multiple users and DL/UL best-effort flows sharing the respective capacities, the above values will correspond to the instantaneous rates, but the effective rates will be decreased. We assume two *independent* systems that follow the Processor-Sharing (PS) scheduling discipline (M/G/1/PS system) [14], [15] for the DL and UL best-effort flows, as shown in Fig. 1. PS is a popular scheduling policy, due to its fairness properties, and is often used to model elastic traffic. Hence, if we assume that the sizes of the best-effort flows (in bits) at location $x$ are drawn independently from two

[2] In the UL scenario, we slightly abuse notation when referring to the SINR, which is of course the SINR at BS i, when the UE transmits with some power $P_{UE}$. Also, the parameters $\zeta_i$, $Z_i$ could be optimized globally or per BS, and could be equal or differentiated. Such an optimization is beyond the scope of this paper, and we'll assume these to be input parameters.

TABLE I. NOTATION

| Variable | Best-Effort Flows | | Dedicated Flows | |
|---|---|---|---|---|
| | Downlink | Uplink | Downlink | Uplink |
| Flow type subscript | 1 | 3 | 2 | 4 |
| Traffic arrival rate (flows/sec) at location $x$ | $\lambda_1(x)$ | $\lambda_3(x)$ | $\lambda_2(x)$ | $\lambda_4(x)$ |
| Flow type probability | $z_b \times z_{\text{DL}}$ | $z_b \times z_{\text{UL}}$ | $z_d \times z_{\text{DL}}$ | $z_d \times z_{\text{UL}}$ |
| Maximum rate (capacity) of the $i$-th BS at location $x$ | $c_{i,1}(x)$ | $c_{i,3}(x)$ | $c_{i,2}(x)$ | $c_{i,4}(x)$ |
| Utilization density of the $i$-th BS at location $x$ | $\varrho_{i,1}(x)$ | $\varrho_{i,3}(x)$ | $\varrho_{i,2}(x)$ | $\varrho_{i,4}(x)$ |
| Total utilization (load) of the $i$-th BS | $0 \le \rho_{i,1} \le 1$ | $0 \le \rho_{i,3} \le 1$ | $0 \le \rho_{i,2} \le 1$ | $0 \le \rho_{i,4} \le 1$ |
| Chance that a specific flow arriving at location $x$ is routed to BS $i$ | $p_{i,1}(x) \in \{0,1\}$ | $p_{i,3}(x) \in \{0,1\}$ | $p_{i,2}(x) \in \{0,1\}$ | $p_{i,4}(x) \in \{0,1\}$ |
| Flow Statistics at location $x$ | Flow-sizes in bits: $y(x)$ | $Y(x)$ | Flow-demand in bps: $E[b(x)]$ | $E[B(x)]$ |

generic distributions, with mean $y(x)$ and $Y(x)$ in the DL and UL, respectively, the corresponding utilization *densities* are

$$\varrho_{i,1}(x) = \frac{\lambda_1(x)}{c_{i,1}(x) \cdot \frac{1}{y(x)}}, \quad \varrho_{i,3}(x) = \frac{\lambda_3(x)}{c_{i,3}(x) \cdot \frac{1}{Y(x)}}. \tag{5}$$

Note that $\varrho_{i,1}(x)$ (resp. $\varrho_{i,3}(x)$) is not the actual utilization of the resources of BS $i$, but rather a measure of the intensity of best-effort traffic demand at location $x$ relative to the available capacity at location $x$, by BS $i$. From the *network-perspective*, these utilizations should not exceed 1 (to avoid congestion) and ideally equalized among different cells (for load-balancing).

### C. Service model for dedicated flows

A dedicated flow is subject to admission control, as it requires some resources for exclusive usage. If a user initiates a new dedicated flow (e.g. an online video game) when the system is already using all its resources, the session will be "blocked". Thus, we chose to apply the M/G/k/k [14] (or, $k$-loss) system, where $k$ expresses the "maximum" number of dedicated flows the BS can serve simultaneously. In queueing terms, this is the number of available *servers* or *resources*.

However, each DL flow might demand a different dedicated rate, so we can approximate the *average* resource constraint $k$ for different BSs at location $x$, as follows [8]. Let there be different types of dedicated flows: a flow of type $i$ requires a data rate of $b_i(x)$ bps, and the ratio of flows with rate $b_i(x)$ is equal to $r_i$ (where $\sum_i r_i = 1$). Then, the average data rate $E[b(x)]$ (in bps) for an incoming DL dedicated flow at $x$ is

$$E[b(x)] = \sum_i b_i(x) r_i. \tag{6}$$

The above is the average DL rate demand. The amount of bandwidth it takes to serve this demand will also depend on the DL physical data rate at location $x$, which is, as we saw,

$$c_{i,2}(x) = (1-\zeta) \cdot w_i \log_2(1 + \text{SINR}_i(x)). \tag{7}$$

Similarly, we can derive the average UL rate demanded $E[B(x)]$, and the UL physical rate $c_{i,4}(x)$. Hence, we can now estimate the maximum number of DL dedicated flows $k_i(x)$ at BS $i$ ($K_i(x)$ in the UL case), as shown in Fig. 1:

$$k_i(x) = \frac{c_{i,2}(x)}{E[b(x)]}, \quad K_i(x) = \frac{c_{i,4}(x)}{E[B(x)]}, \quad i \in \mathcal{B}. \tag{8}$$

$k_i(x)$ can be seen as the number of "servers" (k) for dedicated slots in the above M/G/k/k system at location $x$. As the blocking probability of a $k$-loss system is expected to be monotonically decreasing on $k$, the *user-perspective* wrt the DL dedicated flows corresponds to choose the BS $i$ that offers the maximum $k_i(x)$; whereas wrt the UL the maximum $K_i(x)$.

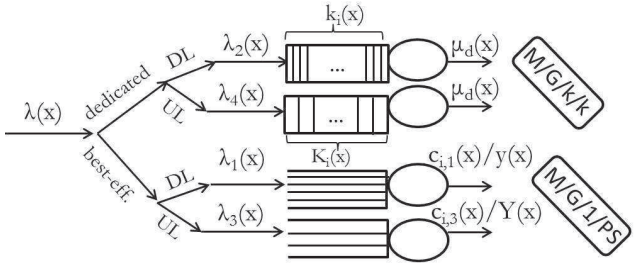



Fig. 1. DL/UL service models for dedicated and best-effort traffic.

Finally, while the above derives the expected amount of demanded resources per flow, each flow might keep its resources for a different duration. Assuming that this random duration comes from a generic distribution with mean $\frac{1}{\mu_d}$, we can define utilization densities for dedicated resources as

$$\varrho_{i,2}(x) = \frac{\lambda_2(x)}{k_i(x)\mu_d(x)} \quad \varrho_{i,4}(x) = \frac{\lambda_4(x)}{K_i(x)\mu_d(x)}, \tag{9}$$

Once more, a network operator would like to keep these values below 1 in the long run, which might lead to redirect a UE to a BS with a smaller $k_i(x)$ for load-balancing purposes.

## III. User Association Problem

We are ready to formulate our optimization problem, and derive the optimal association rules. In doing this, we introduce a routing function $p_{i,y}(x)$, $y \in \{1,2,3,4\}$ that specifies the probability that a flow of type $y$ generated at location $x$ is routed to BS $i$. The association policy consists exactly in finding appropriate values for these routing probabilities. As it will turn out, the optimal values will be either 0 or 1, i.e. the optimal association rule will be deterministic. Before proceeding to the optimization problem, we describe the feasible region for the variables $p_{i,y}(x)$ that is mainly defined by the requirement that the capacity of no BS is exceeded.

*Definition 1:* (Feasibility): The sets $f_y$ of feasible BS loads $\rho_y = (\rho_{1,y}, \rho_{2,y}, ...)$, where $y = 1$ concerns the best-effort utilization and $y = 2$ the dedicated one in the DL, whereas $y = 3$ and $y = 4$ concern them in the UL, are given by

$$\begin{aligned} f_y = \Big\{ \rho_y \mid \rho_{i,y} = \int_{\mathcal{L}} \varrho_{i,y}(x) p_{i,y}(x) dx, \\ 0 \le \rho_{i,y} \le 1 - \epsilon, \\ \sum_{i \in \mathcal{B}} p_{i,y}(x) = 1, \\ 0 \le p_{i,y}(x) \le 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}, \end{aligned} \tag{10}$$

where $\epsilon$ is an arbitrarily small positive constant.

*Lemma 3.1:* The feasible sets $f_1, f_2, f_3, f_4$ are convex.

*Proof:* The proof for the feasible set of DL best-effort flows $f_1$ is presented in [4], and can be easily adapted for the other three cases, as well. ■

Further constraints are introduced, depending on whether a UE at location $x$ is allowed to be attached to different BSs for: (i) different types of flows, (ii) different link scenarios flows.

**(i)** While the offloading of different types of flows to different BSs might be allowed in future setups (e.g. per flow offloading), it is currently not the case. Thus, it should hold $p_{i,1}(x) = p_{i,2}(x)$ and $p_{i,3}(x) = p_{i,4}(x)\ \forall i \in \mathcal{B}$ , i.e. all DL best-effort and dedicated flows should be offloaded to the same BS; similarly in the UL.

**(ii)** Standard setups propose that a UE should be connected to a single BS for both UL and DL traffic, i.e. $p_{i,1}(x) = p_{i,2}(x) = p_{i,3}(x) = p_{i,4}(x)\ \forall i \in \mathcal{B}$ . However, *link-split* [12], allows a UE to offload its UL and DL traffic to different BSs, so $p_{i,1}(x) = p_{i,2}(x)$ and $p_{i,3}(x) = p_{i,4}(x)\ \forall i \in \mathcal{B}$ is sufficient.

### A. Optimal user-association for DL only flows

Let's consider a scenario with both best-effort and dedicated flows, that are all DL (i.e., no UL traffic considered). Following [4], we extend the cost-function to consider performance for dedicated flows as well. The parameters $\{\alpha_1, \alpha_2\}$ control the amount of load-balancing desired for best-effort and dedicated resources, respectively. Parameter $\theta$ reflects which type of traffic is more important. Our cost function is

$$\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}(\rho_{DL}) = \begin{cases} \sum_i \theta \frac{(1-\rho_{i,1})^{1-\alpha_1}}{\alpha_1 - 1} + (1-\theta)\frac{(1-\rho_{i,2})^{1-\alpha_2}}{\alpha_2 - 1}, \text{if } \alpha_1, \alpha_2 \neq 1 \\ \sum_i \theta \log \frac{1}{(1-\rho_{i,1})} + (1-\theta)\log \frac{1}{(1-\rho_{i,2})}, \text{if } \alpha_1 = \alpha_2 = 1 \\ \sum_i \theta \frac{[(1-\rho_{i,1})]^{1-\alpha_1}}{\alpha_1 - 1} + (1-\theta)\log \frac{1}{(1-\rho_{i,2})}, \text{ if } \alpha_2 = 1 \\ \sum_i \theta \log \frac{1}{(1-\rho_{i,1})} + (1-\theta)\frac{(1-\rho_{i,2})^{1-\alpha_2}}{\alpha_2 - 1}, \text{ if } \alpha_1 = 1. \end{cases} \tag{11}$$

where $\rho_{DL} = \left[\rho_1; \rho_2\right]$ and $\mathcal{F}_{DL} = \left[f_1; f_2\right]$ depict the BS loads in the respective dimensions and their feasible values.

*Theorem 3.2:* If the feasible domain $\mathcal{F}_{DL}$ of the problem

$$\min_{\rho_{DL}} \left\{ \phi^{(DL)}_{\alpha_1,\alpha_2,\theta}(\rho_{DL}) | \rho_{DL} \in \mathcal{F}_{DL} \right\} \tag{12}$$

is non-empty, the optimal user-association rule wrt the additional constraint $p_{i,1}(x) = p_{i,2}(x)$ discussed earlier, is

$$i(x) = \arg\max_{i \in \mathcal{B}} \frac{(1-\rho^*_{i,1})^{\alpha_1}(1-\rho^*_{i,2})^{\alpha_2}}{e_1(x)(1-\rho^*_{i,2})^{\alpha_2} + e_2(x)(1-\rho^*_{i,1})^{\alpha_1}}, \tag{13}$$

where $\rho^*_{DL} = [\rho^*_1; \rho^*_2]$ is the optimal load vector (the solution to Problem (12)), $e_1(x) = \frac{\theta y(x) z_{DL} z_b}{c_{i,1}(x)}$ and $e_2(x) = \frac{(1-\theta) z_{DL} z_d}{\mu(x) k_i(x)}$.

*Proof:* The proof is presented in the Appendix. ■

"User vs. Network perspective (Fairness)". $\alpha_1$ and $\alpha_2$ are the parameters that trade off user related performance for network-related one. Detailed discussion for the below claims regarding $\alpha_1$ can be found in [4], whereas regarding $\alpha_2$ in the Appendix.

- *User-perspective:* $\alpha_1 = 0$ maximizes the average physical rate for the best-effort flows as defined in Eq. (4), whereas $\alpha_2 = 0$ maximizes the average dedicated servers for dedicated flows as defined in Eq. (8).
- *Optimizing related QoS metrics:* if $\alpha_1 = 2$, the average delay is minimized, since the cost function for best effort flows becomes equal to the expected delay in an M/G/1/PS system. If $\alpha_2 = 1$ the corresponding optimal rule becomes equivalent to the average *idle* dedicated servers in a k-Loss system, so the actual blocking probability for dedicated flows is minimized.
- *Network Perspective:* As $\alpha_1 \to \infty$, we minimize the maximum BS utilization, i.e. load balancing between the $\rho_1$ is achieved. Similar for $\alpha_2$ and $\rho_2$'s.

"Dedicated vs. best-effort flows performance". Parameter $0 \leq \theta \leq 1$ is a linear weight factor deciding the importance of optimizing dedicated flow ($\theta \to 0$) vs. best effort flow performance ($\theta \to 1$). Different operators might choose different values at different times of day, service level agreements etc.

### B. Optimal user-association for UL only flows

The case of UL traffic only is entirely symmetrical to the DL case just addressed. For completeness, we state the cost function and optimal association rule.

$$\phi^{(UL)}_{\alpha_3,\alpha_4,\Theta}(\rho_{UL}) = \begin{cases} \sum_i \Theta \frac{(1-\rho_{i,3})^{1-\alpha_3}}{\alpha_3 - 1} + (1-\Theta)\frac{(1-\rho_{i,4})^{1-\alpha_4}}{\alpha_4 - 1}, \text{if } \alpha_3, \alpha_4 \neq 1 \\ \sum_i \Theta \log \frac{1}{(1-\rho_{i,3})} + (1-\Theta)\log \frac{1}{(1-\rho_{i,4})}, \text{if } \alpha_3 = \alpha_4 = 1 \\ \sum_i \Theta \frac{(1-\rho_{i,3})^{1-\alpha_3}}{\alpha_3 - 1} + (1-\Theta)\log \frac{1}{(1-\rho_{i,4})}, \text{if } \alpha_4 = 1 \\ \sum_i \Theta \log \frac{1}{(1-\rho_{i,3})} + (1-\Theta)\frac{(1-\rho_{i,4})^{1-\alpha_4}}{\alpha_4 - 1}, \text{if } \alpha_3 = 1. \end{cases} \tag{14}$$

If $e_3(x) = \frac{\Theta Y(x) z_{UL} z_b}{c_{i,3}(x)}$, $e_4(x) = \frac{(1-\Theta) z_{UL} z_d}{\mu(x) K_i(x)}$, our rule becomes

$$i(x) = \arg\max_{i \in \mathcal{B}} \frac{(1-\rho^*_{i,3})^{\alpha_3}(1-\rho^*_{i,4})^{\alpha_4}}{e_3(x)(1-\rho^*_{i,4})^{\alpha_4} + e_4(x)(1-\rho^*_{i,3})^{\alpha_3}}. \tag{15}$$

### C. Optimal DL and UL user-association

We are now ready to consider *jointly* the DL and UL performance while deciding the optimal user-associations. As already mentioned we are going to investigate the optimal association rules either when UL/DL split is offered, or not.

*1) **"Split Scenario"**: Link split* (or DL/UL decoupling) allows each UE to be associated with two BSs for its DL and UL offloading [12], [13], to maximize systems performance in both dimensions. So, the independent DL and UL associations can be found by *separately* solving the optimization problems described in Sections III-A, III-B, respectively.

*2) **"Non-Split Scenario"**:* However, depending on the operator's capabilities, the link-split might *not* be applicable. Hence, if $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ we form a new cost function to optimally associate each UE with only one BS wrt to $\tau$:

$$\phi_{\alpha,\theta,\Theta}(\rho) = \tau \phi^{(DL)}_{\alpha_1,\alpha_2,\theta}(\rho_{DL}) + (1-\tau)\phi^{(UL)}_{\alpha_3,\alpha_4,\Theta}(\rho_{UL}) \tag{16}$$

where $\rho = \left[\rho_1; \rho_2; \rho_3; \rho_4\right]$ and $\mathcal{F} = \left[f_1; f_2; f_3; f_4\right]$.

"DL vs. UL performance". $0 \leq \tau \leq 1$ is a linear weight factor deciding the importance of optimizing the DL performance ($\tau \to 1$) vs. the UL performance ($\tau \to 0$).

*Theorem 3.3:* If the feasible domain $\mathcal{F}$ of the problem

$$\min_{\rho} \left\{ \phi_{\alpha,\theta,\Theta}(\rho) | \rho \in \mathcal{F} \right\} \tag{17}$$

is non-empty, the optimal user-association rule wrt the additional constraint $p_{i,1}(x) = p_{i,2}(x) = p_{i,3}(x) = p_{i,4}(x)$, is

$$i(x) = \arg\max_{i\epsilon\mathcal{B}} \frac{\prod_{l=1}^{4}((1-\rho_{i,l}^{*})^{\alpha_l})}{\sum_{j=1}^{4} e_j(x)\prod_{l=1,l\neq j}^{4}((1-\rho_{i,l}^{*})^{\alpha_l})}, \quad (18)$$

where $\rho^* = [\rho_1^*; \rho_2^*; \rho_3^*; \rho_4^*]$ is the optimal load vector, $e_1(x) = \tau\frac{\theta y(x) z_{DL} z_b}{c_{i,1}(x)}$, $e_2(x) = \tau\frac{(1-\theta) z_{DL} z_d}{\mu(x) k_i(x)}$, $e_3(x) = (1-\tau)\frac{\Theta Y(x) z_{UL} z_b}{c_{i,3}(x)}$ and $e_4(x) = (1-\tau)\frac{(1-\Theta) z_{UL} z_d}{\mu(x) K_i(x)}$.

*Proof:* The proof follows very similar steps to the proof of Theorem 3.2 in the Appendix. Due to space limitations, we refer the interested reader to [16] for the detailed proof. ■

*Remark 1:* It is interesting to look a bit deeper into the above optimal association rule. When considering multiple conflicting objectives, it is optimal to associate a user to the BS that maximizes the *harmonic mean* of the individual association rules when considering each objective alone. E.g., assume a simple scenario with only UL and DL best-effort traffic. And assume the following BS options for a user: (BS A) gives 50Mbps DL and only 1Mbps UL; (BS B) 200Mbps DL and 0.5Mbps UL; (BS C) 20Mbps DL and 5Mbps UL. If we care about UL and DL traffic equally (i.e. $\tau = 0.5$), one might assume that the BS with the highest sum (or arithmetic average) of rates should be chosen (i.e. BS B). However, the optimal BS, according to our rule is the BS that maximizes the harmonic mean, namely BS C. Harmonic means appear in a number of physical examples, such as parallel resistances, where the total resistance is the harmonic mean of the parallel ones. In that case, increasing the total resistance would require increasing the smallest resistance. In that sense, the optimal rule can be seen as applying a max-min principle among the various conflicting objectives. However, a number of systems parameters (e.g, $\theta$, $\tau$, etc.) also enter the policy rule, changing the weights of each "branch" on this harmonic mean. Note that, the harmonic mean usage further allows to add more dimensions in our setup and flexibly derive the optimal rule.

## IV. SDN-BASED IMPLEMENTATION

Now, we propose an online centralized algorithm that achieves global optimum in an iterative manner. This algorithm takes as inputs (i) the overall network status, and (ii) some high level system-parameters, in order to identify the optimal associations. This procedure is rather facilitated from a SDN architecture that offers a centralized programmable control for the underlying network. Following the SDN outline, we consider four planes as illustrated in Fig. 2:

**Application tier:** The operator determines some system-related parameters (e.g. $\alpha, \theta, \tau, \zeta$ etc), and advertises them to the controller, at the end of the $k$-th iteration period.
**Controller tier:** At each $k$ period, the controller receives (i) the above-mentioned system-related parameters, and (ii) some network-related parameters (e.g. $z_{DL}$, size-files etc) as well as the 4-dimensional load vector $\rho^{(k)}$ from the application and network tier, respectively. Then, it determines and advertises to BSs the optimal associations derived from Eq. (13,15,18).
**Network tier:** Each $k$-th period, BSs either apply or indicate to users the optimal rules depending on how the association is managed in the network. At the end of $k$, they measure and advertise to the controller their average load levels $\rho^{(k)}$, and the network-relater parameters (e.g. $z_{DL}$, size-files etc).
**User tier:** At each $k$-th period, a UE at location $x$ is associated or triggers the association procedure to the new BSs.

TABLE II. SIMULATION PARAMETERS

| Parameter | Variable | Value |
|---|---|---|
| Transm. Power of Macro BS/ SC/ UE | $P_{eNB}/P_{SC}/P_{UE}$ | 43/24/12 dBm |
| System Bandwidth for DL, UL | $w/W$ | 10/10 MHz |
| Noise Power Density | $N_0$ | -174 dBm/Hz |
| Splitting parameter for DL, UL | $\zeta_i, Z_i$ | 0.5/0.5 |
| Average sizes: DL/UL best-eff. flows | $y(X)/Y(x)$ | 100/20 Kbytes |
| Average traffic demands: DL/UL ded. flows | $E[b(x)]/E[B(x)]$ | 512, 256 kbps |
| Ratio of best-effort, ratio of DL flows | $z_b, z_{DL}$ | 0.3,0.6 |

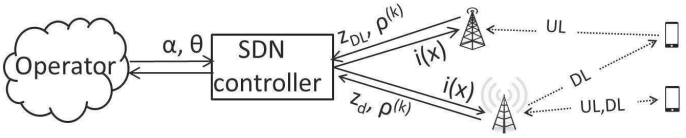



Fig. 2. Applicability to the SDN architecture.

The above iteration provably converges to the global optimal point with a simple modification of the proof in [4].

## V. SIMULATIONS

In this section we briefly present some numerical results and discuss related insights. We consider a $2 \times 2$ $km^2$ area. Fig. 3(a) shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ ($flows/hour$ per unit area) (blue implying low traffic and red high), with 3 hotspots. Furthermore, we assume that this area is covered by four macro BSs (shown with asterisks) and six SCs (shown with triangles) as depicted in Fig. 3(b)-(c), and Fig. 4(a)-(b). We also consider standard parameters as adopted in 3GPP [17], listed in Table II[3]. Throughout this section, if not explicitly mentioned, we assume the *"Non-Split Scenario"*, described in III-C2, and $\theta = \Theta = 0.5$.

Before proceeding, we setup a metric that reflects the "network-related" performance goal, namely load balancing. Hence, we introduce the Mean Squared Error (MSE) between the utilization of different BSs, normalized to 1:

$$\text{MSE}_1 = \frac{1}{2 * h}\sum_i\sum_j(\rho_{i,1} - \rho_{j,1})^2, \quad (19)$$

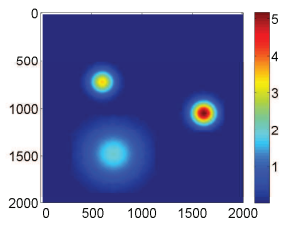
where $h$ is the normalizing factor and is equal to $h = \left\lfloor\frac{N}{2}\right\rfloor \times \left\lceil\frac{N}{2}\right\rceil$, and $N$ the total number of BSs. The higher the $\text{MSE}_1$, the more imbalanced the load for the DL "best-effort" resources across different BSs. We can define similar metrics $MSE_2, MSE_3, MSE_4$ for the remaining utilization values.

*User vs. Network Perspective* trade-off is considered qualitatively in Fig. 3(b)-3(c) along with Fig. 3(e) that quantifies via a Table some performance metrics, wrt the DL scenario (Section III-A). Concerning the *user-perspective*, Fig. 3(b) outlines the optimal associations if $\alpha_1 = \alpha_2 = 0$. Thus, each UE is attached to the BS that offers the *highest DL SINR* and promises higher average DL physical rate for best effort flows $E[c_1]$, and more average "dedicated" servers $E[k]$; i.e. most of UEs are attached to macro BSs due to their

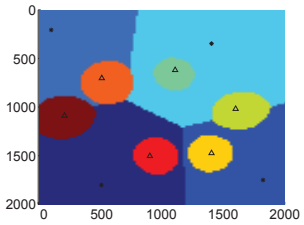
[3]As for (i) the sizes and ratios of different flows, (ii) splitting parameters, we can use different values in order to capture different simulation scenarios.

high power transmission, and fewer to SCs, forming small circles around them. Consequently, macrocells are overloaded and load imbalance within the cells is sharpened (increased $MSE_1, MSE_2$; see line 1 of Fig. 3(e)). However, in Fig. 3(c) we emphasize the *network-perspective* (load-balancing) and set $\alpha_1 = \alpha_2 = 10$. Now, (i) SCs increase their coverage area in order to offload the overloaded macro BSs, (ii) *all* the "heavily" loaded (due to the hotspots) BSs, shrink their coverage. Thus load imbalance is alleviated, at the cost of $E[c_1], E[k_1]$ (see line 2 of Fig. 3(e)).

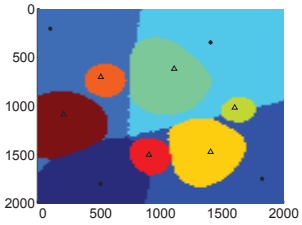
Although in the previous scenarios the best-effort- and dedicated- related traffic rules (represented from $\alpha_1, \alpha_2$) are aligned, one could ask how would two conflicting optimization objectives affect our network? The answer lays in the usage of $\theta$, that judges which objective carries more importance. E.g., an operator has two main goals: (i) to maximize the user QoS for "dedicated" traffic captured by $E[k]$ (set $\alpha_2 = 0$), (ii) to better balance the utilization of best-effort resources between BSs (set $\alpha_1 = 10$). As shown in Fig. 3(d), if $\theta \to 0$ $E[k]$ is maximized, whereas as $\theta \to 1$ $MSE_1$ is minimized, and each objective comes at the price of the other.



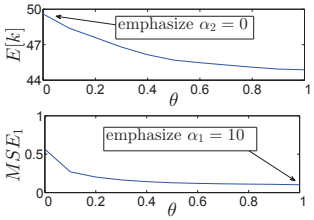

(a) Traffic arrival rate.




(b) Pure user-persp. $\alpha_1 = \alpha_2 = 0$.




(c) Pure netw.persp. $\alpha_1 = \alpha_2 = 10$.




(d) Ded. vs. Best-effort flows.

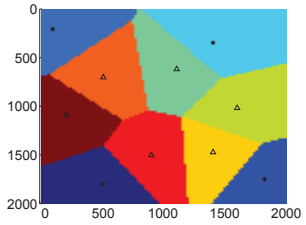
| | User perspective | | Network perspective | |
|---|---|---|---|---|
| | $E[c_1]$ (Mbps) | $E[k]$ | $MSE_1$ | $MSE_2$ |
| Fig. 3(b) $\alpha_{\{1,2\}} = 0$ | 24.3 | 48 | 0.31 | 0.27 |
| Fig. 3(c) $\alpha_{\{1,2\}} = 10$ | 21.3 | 42 | 0.003 | 0.002 |

(e) Quantitative considerations of different trade-offs.

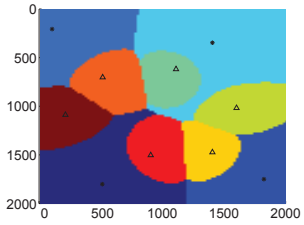
Fig. 3. Optimal user-associations wrt the $\alpha_1, \alpha_2, \theta$, in the DL only scenario.

*DL vs. UL performance* (Section III-C) is considered in Fig. 3(b), 4(a)-4(b). These figures are supplemented by specific performance metrics of user performance shown in Fig. 4(c). We remind to the reader that parameter $\tau$ trades-off between the UL and DL performance. As already discussed, Fig. 3(b) outlines the optimal associations if the whole emphasis is on the *DL performance* ($\tau = 1$): however this hurts the UL performance due to the asymmetric transmission powers of the UEs and BSs (see line 1 of Fig. 4(c)). In Fig. 4(a) we move the emphasis on the *UL performance* ($\tau = 0$), and each UE is attached to the nearest BS, in order to minimize the path loss [13] and enhance the UL performance; this hurts its DL performance though (see line 3 of Fig. 4(c)). Finally, Fig. 4(b) shows the optimal coverage areas when one assigns equal importance to the UL and DL performance (i.e. $\tau = 0.5$): this moderates both DL and UL performance (line 2 of Fig. 4(c)).

Hence, in the above-mentioned *Non-Split Scenario* is impossible to achieve optimal UL/DL performance *simultaneously*. Using $\tau$, we can trade-off which carries more importance while selecting a single BS for association, though. But, according to the "*Split Scenario*" each UE is attached to two BSs: one that maximizes its DL, and one that maximizes its UL performance, as shown in Fig. 3(b), 4(a), respectively.




(a) UL optimal $\tau = 0$.




(b) Equal emphasis $\tau = 0.5$.

| | DL performance | | UL performance | |
|---|---|---|---|---|
| | $E[c_1]$ (Mbps) | $E[k]$ | $E[c_3]$ (Mbps) | $E[K]$ |
| Fig. 3(b), $\tau = 1$ | 24.3 | 48 | 6.5 | 25 |
| Fig. 4(b), $\tau = 0.5$ | 22.9 | 45 | 7.5 | 29 |
| Fig. 4(a), $\tau = 0$ | 18.1 | 36 | 8.5 | 33 |

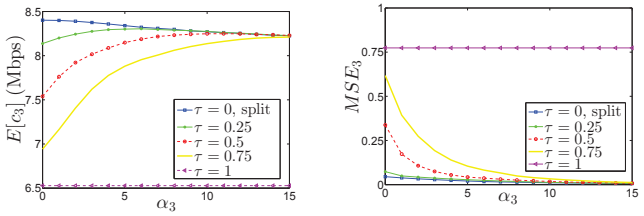
(c) Quantitative considerations of different trade-offs.

Fig. 4. Optimal user-associations wrt $\tau$, for the pure user-persp. scenario.

Fig. 5 illustrates the impact of the user vs. network perspective trade-off wrt different $\tau$. Firstly, *given* $\tau = 0$: $\alpha_3 = 0$ maximizes $E[c_3]$, and as $\alpha_3$ is increased the network-perspective is enhanced (lower $MSE_3$) at the price of user-perspective (decreased $E[c_3]$); this is consistent with our theory. However, *for higher values of* $\tau > 0$: $\alpha_3 = 0$ *does not* maximize $E[c_3]$. This is reasonable since the emphasis put in the DL ($\tau > 0$) will attempt to enhance the DL performance, so associate the users with the macro BSs (higher power transmission), and eventually: (i) overload BSs UL resources, (ii) weaken the user UL rates. Interestingly, as we increase $\alpha_3$ (move to load balancing) more users are associated with the SCs to offload the macro BSs, consequently $E[c_3]$ is enhanced, up to the point that is maximized and starts obeying the general decreasing rule. This constitutes a very interesting trade-off since it suggests that we can achieve *both* enhanced user and network perspective, in different scenarios. Summing up, in the above-mentioned *Non-Split Scenario* the UL performance is bounded in terms of $E[c_3]$ and $MSE_3$ (the bounds are assured *only* when $\tau = 0$); *Split Scenario* though, always guarantees these optimal bounds, as depicted in Fig. 5.

## VI. Conclusion

In this paper, we considered the user-association problem for future dense HetNets. We propose a theoretical framework that optimally assigns users to BSs, considering: (i) user-centric and network-centric performance goals, (ii) traffic consisting of both best-effort and dedicated flows, and (iii) UL and DL performance. Optimal association rules are analytically derived along with our framework, and initial simulation results are presented that reveal interesting tradeoffs.




(a) $E(c_3)$ vs. $\alpha_3$. (b) $MSE_3$ vs. $\alpha_3$.

Fig. 5. Trade-off: "user" vs "network" perspective in the UL scenario.

## REFERENCES


[1] *LTE Release 10 & beyond*, 3GPP Long Term Evolution, 2011.

[2] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Communications*, 2014.

[3] *Backhaul for urban small cells: a topic brief*, Small Cell Forum, 2014.

[4] 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*, 2012.

[5] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. on Selected Areas in Communications*, 2011.

[6] S. Liu and J. Virtamo, "Inter-cell coordination with inhomogeneous traffic distribution," in *Proc. IEEE Next G. Int. Design and Eng.*, 2006.

[7] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Comm. Surveys*, 2013.

[8] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *Proc. IEEE Globecom*, 2014.

[9] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[10] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Joint uplink and downlink cell selection in cognitive small cell heterogeneous networks," in *Proc. IEEE Globecom*, 2014.

[11] P. Rost, A. Maeder, and X. Perez-Costa, "Asymmetric uplink-downlink assignment for energy-efficient mobile communication systems," in *Vehicular Technology Conference (VTC Spring)*, 2012.

[12] G. T. . v.12.0.0 Rel.12, *Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects*. Academic press, 2013.

[13] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Globecom*, 2014.

[14] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Imperial college press, 2010.

[15] A. Elwalid and D. Mitro, "Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes," in *Proc. IEEE Infocom*, 1999.

[16] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal downlink-uplink user association in HetNets with traffic differentiation." Tech. Report RR-15-301, Eurecom, 2015.

[17] *3GPP, Technical Report LTE; Evolved Universal Terrestrial Radio Access (E-UTRA)*, TR 136 931, 2011.


## APPENDIX

### A. Impact of $\alpha_2$ on the user and network perspective

To better elucidate the problem at hand we assume $\theta \to 0$ (emphasize the dedicated flows performance). (13) becomes[4]

$$i(x) = \arg\max_{j\in\mathcal{B}} k_j(x)[(1-\rho^*_{j,2})]^{\alpha_2}. \tag{20}$$

From (20), if $\alpha_2 = 1$ the average number of *idle* "servers" $k_{j,1}(x)\cdot(1-\rho^*_{j,2})$ is maximized. Equivalently, the actual blocking probability is minimized, since it is monotonically decreasing on the *idle* servers [14]. Similarly, if $\alpha_2 = 0$ or $\alpha_2 \to \infty$, the optimal rule implies maximization of $k$ and load balancing (minimization of the maximum utilization), respectively.

### B. Proof of Eq. (13)

We prove that (13) is the optimal association rule of Problem (12), subject to the additional constraint $p_{i,1}(x) = p_{i,2}(x)$. Let $\rho^* = [\rho^*_1; \rho^*_2]$ be the optimal solution of Problem (12). Problem (12) is a convex optimization because its feasible set $\mathcal{F} = [f_1; f_2]$ has been proved to be convex in Lemma 3.1, and the objective function $\phi^{(DL)}_{\alpha,\theta}(\rho)$ is also convex (due to the summation and linear combinations of the convex function $\phi_\alpha(\rho)$ that is proven to be convex in [4]). Hence, it is adequate to check the following condition for optimality

$$\langle \nabla\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}(\rho^*), \Delta\rho^*\rangle \geq 0 \tag{21}$$

for all $\rho \in \mathcal{F}$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for $\rho$ and $\rho^*$, respectively. Using the deterministic cell coverage generated by(13), the optimal association rule is given by:

$$p^*_i(x) = \mathbf{1}\left\{ i = \arg\max_{j\in\mathcal{B}} \frac{(1-\rho^*_{j,1})^{\alpha_1}(1-\rho^*_{j,2})^{\alpha_2}}{e_1(x)(1-\rho^*_{j,2})^{\alpha_2} + e_2(x)(1-\rho^*_{j,1})^{\alpha_1}} \right\}. \tag{22}$$

Then the inner product in Eq. (21) can be written as:

$$\begin{aligned}
&\langle \nabla\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}(\rho^*), \Delta\rho^*\rangle = \sum_{z=1}^{2} \frac{\partial\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}}{\partial\rho_z}(\rho^*)(\rho_z - \rho^*_z) \\
&= \frac{\partial\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}}{\partial\rho_1}(\rho^*)(\rho_1-\rho^*_1) + \frac{\partial\phi^{(DL)}_{\alpha_1,\alpha_2,\theta}}{\partial\rho_2}(\rho^*)(\rho_2-\rho^*_2) \\
&= \theta\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_{i,1})^{\alpha_1}}(\rho_{i,1}-\rho^*_{i,1}) + (1-\theta)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_{i,2})^{\alpha_2}}(\rho_{i,2}-\rho^*_{i,2}) \\
&= \sum_{i\in\mathcal{B}} \frac{\theta\int_L \varrho_{i,1}(x)(p_i(x)-p^*_i(x))dx}{(1-\rho_{i,1})^{\alpha_1}} + \frac{(1-\theta)\int_L \varrho_{i,2}(x)(p_i(x)-p^*_i(x))dx}{(1-\rho_{i,2})^{\alpha_2}} \\
&= \int_L \lambda(x)\sum_{i\in\mathcal{B}}(p_i(x)-p^*_i(x))\Big[\frac{e_1(x)(1-\rho^*_{i,2})^{\alpha_2} + e_2(x)(1-\rho^*_{i,1})^{\alpha_1}}{(1-\rho^*_{i,1})^{\alpha_1}(1-\rho^*_{i,2})^{\alpha_2}}\Big]dx,
\end{aligned} \tag{23}$$

where $e_1(x) = \frac{\theta y(x) z_{DL} z_b}{c_{i,1}(x)}$ and $e_2(x) = \frac{(1-\theta) z_{DL} z_d}{\mu(x) k_i(x)}$. Note that,

$$\begin{aligned}
&\sum_{i\in\mathcal{B}} p_i(x)\frac{e_1(x)(1-\rho^*_{i,2})^{\alpha_2} + e_2(x)(1-\rho^*_{i,1})^{\alpha_1}}{(1-\rho^*_{i,1})^{\alpha_1}(1-\rho^*_{i,2})^{\alpha_2}} \geq \\
&\sum_{i\in\mathcal{B}} p^*_i(x)\frac{e_1(x)(1-\rho^*_{i,2})^{\alpha_2} + e_2(x)(1-\rho^*_{i,1})^{\alpha_1}}{(1-\rho^*_{i,1})^{\alpha_1}(1-\rho^*_{i,2})^{\alpha_2}}
\end{aligned} \tag{24}$$

holds because $p^*(x)$ in (22) is an indicator for the minimizer of $\frac{e_1(x)(1-\rho^*_{i,2})^{\alpha_2} + e_2(x)(1-\rho^*_{i,1})^{\alpha_1}}{(1-\rho^*_{i,1})^{\alpha_1}(1-\rho^*_{i,2})^{\alpha_2}}$. Hence, (21) holds.

[4] One can also *analytically* derive the same association rule, by assuming $\theta \to 0$ directly in the cost function (11), and further minimize it.