Time and Location Aware Mobile Data Pricing

Qian Ma*, Ya-Feng Liu†, and Jianwei Huang*
*Dept. of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China.
†LSEC, AMSS, Chinese Academy of Sciences, Beijing, China.

***Abstract*—Mobile users' social behaviors often lead to significant temporal and spatial variations of mobile traffic. This could create severe cellular network congestion in peak hours and hot spots. This paper presents an initial study on designing the time and location aware pricing scheme to incentivize users to smooth traffic and reduce network congestion. We derive the optimal pricing scheme through analyzing a two-stage decision process, where the operator announces the time and location aware prices in Stage I, and users schedule their mobile traffic accordingly in Stage II. We can translate such a two-stage decision problem into a bilevel optimization problem, which is NP-hard and challenging to solve. We propose an easily implementable algorithm, which utilizes a penalty method and a block coordinate decent algorithm to solve the problem. The resultant pricing scheme ensures a win-win situation for both the operator and users. Our simulation shows that the operator can reduce the extra cost for provisioning the peak traffic by up to 98.70%, and users can increase their total payoff by up to 106.10%, comparing with a time and location independent pricing benchmark.**

I. Introduction

Cisco has predicated that the global mobile data demand will grow at an annual rate of 66% for the next few years, but the mobile cellular network capacity will only grow with an annual rate of 29% [1], [2]. As a result, the total mobile data demand may surpass the total network capacity globally as early as in 2014 [3]. In order to alleviate the tension between supply and demand, the cellular operators have been trying to increase the network capacity through adopting new communication technologies (such as shifting from 3G to 4G technologies) and obtaining more spectrum (such as utilizing the TV white space for cellular communications [4]). Another equally effective approach is to use pricing to shape the traffic to better fit into the existing network capacity [5].

One widely used pricing strategy for shaping cellular data traffic is the usage-based pricing. For example, AT&T in the USA has adopted a tiered usage-based monthly pricing plan since 2010 [6]. Although the usage-based pricing is simple to implement, it is not always the most effective approach, as it ignores the stochastic nature of traffic over time and location.

From the cellular operator's point of view, the aggregate mobile data traffic varies significantly with time and location, and there are easily identifiable peak hours and crowded locations (such as business hours at commercial buildings and night time in highly populated residential areas) [7]. In fact, a major cost for the cellular operator is to cope with the peak demands at certain time slots and locations; meanwhile, the network capacity is not fully utilized at other time slots and locations. If a pricing scheme is aware of such traffic stochastics, and provides incentives for users to schedule the transmissions properly so as to smooth the traffic, it will lead to a win-win situation for both the operator and users.

Time and location aware pricing is not something completely new in the industry. Some heuristic schemes of this type have already existed in practice, such as MTN's dynamic tariffing in Africa for pricing voice calls [9]. The effectiveness of these existing practices, together with the exploding wireless data traffic, motivates us to provide a rigorous holistic design for time and location aware pricing for wireless data traffic. Notice that cellular operators usually charge data traffic based on volume and charge voice calls based on call durations, hence the optimal pricing schemes for these two types of traffic will be very different.

Research results regarding time-aware (but location independent) pricing for mobile data traffic only emerged very recently. Joe-Wong *et al.* in [10] demonstrated the effectiveness of time-aware pricing in terms of encouraging users to shift traffic to later non-peak hours. Tadrous *et al.* in [11] illustrated the possibility to use time-aware pricing to encourage users to pre-download data before peak hours. However, neither results exploited the spatial dynamics of the traffic.

The only result regarding location-aware data pricing is the experiments from AT&T [12]. This pricing scheme separates the whole network area into several regions, and optimizes the pricing for each region independently. Law in [12] demonstrated that a location-based pricing can reduce network congestion, but the study did not consider users' mobilities, the impact of pricing on users' payoffs, and the time dimension traffic variations.

The goal of our study is to design a time and location aware pricing scheme to provide benefits to both the cellular operator and mobile users. Our main results and contributions are summarized as follows.

- To the best of our knowledge, this is the first result regarding a holistic optimal design of the mobile data pricing jointly in the time and spatial dimensions.
- We capture the interactions between the cellular operator and users as a two-stage decision process, which considers users' global and local mobility patterns in the spatial domain and users' delay preference in the time domain.

This work is supported in part by the General Research Funds (Project Number CUHK 412713 and CUHK 412511) established under the University Grant Committee of the Hong Kong Special Administrative Region, China, and in part by the National Natural Science Foundation of China, Grants 11331012 and 11301516.

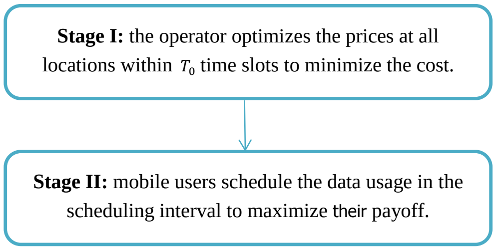

Figure 1: Two-Stage Decision Process

- We show that our model leads to a bilevel pricing optimization problem which is NP-hard, and we propose an easily implementable algorithm combining a penalty method and the block coordinate decent algorithm to solve the problem.
- Simulations show that both the cellular operator and users benefit from the time and location aware pricing scheme. The cellular operator reduces the extra cost for provisioning the peak traffic by up to 98.70%, and users increase their payoffs by up to 106.10%, comparing with a time and location independent pricing benchmark.

The rest of the paper is organized as below. We introduce the two-stage decision model in Section II. In Sections III and IV, we analyze users' decision problems in Stage II and the cellular operator's pricing decision in Stage I, respectively. In Section V, we show that the two-stage optimization problem can be reformulated as a single bilevel optimization problem, and propose a penalty method and the block coordinate descent (BCD) algorithm to solve the problem. We verify the effectiveness of the proposed pricing scheme in Section VI. We finally conclude in Section VII.

II. System Model

We consider a cellular network where the cellular operator determines prices first, then users decide how to use the data services based on the prices. We assume that users are price-takers, who do not anticipate the impact of their demands on the operator's prices. Such a price-taking behavior is reasonable, as there are usually many users subscribing for the service from the same operator, and the impact of a single user on the entire network is negligible.

We capture the above sequential interactions as a two-stage decision process. During Stage I, the operator announces the prices over different time slots (e.g., different hours) and locations (e.g., different base stations). In Stage II, each user decides his demand in each time slot based on the prices and his own mobility. Figure 1 shows the two-stage decision process. We notice that the current time and location independent usage based pricing scheme (such as the one used by AT&T) is a special case of our general model.

In the time domain, the cellular operator makes the pricing decisions in Stage I, for an entire period of $\mathcal{T}_0 = \{1, 2, \ldots, T_0\}$ time slots. In Stage II, when a mobile user becomes active in time $t \in \mathcal{T}_0$ and wants to consume certain amount of traffic[1], he can schedule the actual data consumption during one or more of the following time slots: $\{t, t+1, \ldots, T_t\}$, where $T_t = \min\{t+T-1, T_0\}$. We call T the scheduling interval, which is usually smaller than the operator's pricing span T_0. Figure 2 shows an example with $T_0 = 12$ and $T = 4$.

[1]For simplicity, we consider the downlink transmission in this paper, where the base station sends traffic to users. The uplink transmission can be analyzed similarly.

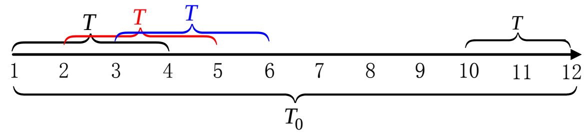

Figure 2: One example for $T_0 = 12$ and $T = 4$. A user becomes active at time slot 3 can schedule in time slots $3, 4, 5$ and 6, but a user who becomes active at time slot 10 can only schedule his traffic demand in time slots $10, 11$ and 12.

In the spatial domain, we take into consideration users' mobility patterns that capture their sociological orbits [8]. The cellular operator is able to construct aggregate *mobility profiles* for the entire user population based on historical observations. There are two types of mobility profiles: the global (long-term) one which captures users' mobility at a large time scale (say over a whole day or even a whole week, which is described as T_0 time slots in this paper), and the local (short-term) one which captures users' mobility at a smaller time scale (say over several consecutive time slots).

We denote the set of possible locations as $\mathcal{L} = \{1, 2, \ldots, L\}$. The global mobility profile is:

$$\boldsymbol{\alpha} = \{\alpha_{t,l} : \alpha_{t,l} \geq 0, \sum_{l=1}^{L} \alpha_{t,l} = 1, \forall t \in \mathcal{T}_0, l \in \mathcal{L}\}, \quad (1)$$

where $\alpha_{t,l}$ represents the probability of a user appearing at location l at time slot t from a long-term point of view. The operator can make more precise predications for the users' mobility across time slots and locations through the local mobility profile:

$$\boldsymbol{\beta^a} = \{\beta^a_{t,l}(t', l') : \beta^a_{t,l}(t', l') \geq 0, \sum_{l'=1}^{L} \beta^a_{t,l}(t', l') = 1, \forall t \in \mathcal{T}_0, l \in \mathcal{L}, t' \in \mathcal{T}_t, l' \in \mathcal{L}\}. \quad (2)$$

Here $\mathcal{T}_t = \{t+1, t+2, \ldots, T_t\}$, $\beta^a_{t,l}(t', l')$ represents the probability of a user appearing at location l' at time slot t' given that he has appeared at location l at time slot t, and a denotes the user type, which will be explained in details in Section III. Reference [13] provides more detailed discussions regarding how to construct the mobility profiles through learning users' movement history. In this paper, we will assume that the mobility profiles $\boldsymbol{\alpha}$ and $\boldsymbol{\beta^a}$ are given system parameters.

Since the cellular operator will make price decisions based on users' scheduling decisions, we will analyze the two-stage decision process through backward induction.

III. Users' Decision in Stage II

In Stage II, a user needs to schedule his data usage to maximize his payoff (i.e., utility minus payment), given the fixed prices announced by the operator in Stage I:

$$\boldsymbol{p} = \{p(t, l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}\}. \quad (3)$$

Here $p(t, l)$ is the price per unit of data traffic at time t and location l.

A user's utility depends on several factors, including a *utility function* that characterizes the user's satisfaction level of consuming certain amount of traffic in a single time slot, the *delay tolerance parameter* $\delta \in [0,1]$ that captures the user's willingness to wait, and the user's *mobility pattern* which predicts his relationship between time and location in the next T time slots. We will divide the user population into a set $\mathcal{A} = \{1, 2, \ldots, A\}$ of types, where each type of users have the same utility function, delay tolerance parameter, and mobility profile.

As a concrete example, consider a type a user who becomes active at time t and location l with a demand of $x_0^a(t,l)$ data traffic. The subscript 0 represents that this is the "initial" demand before scheduling. If the prices announced by the operator in Stage I are time and location *independent*, then it is easy to show that the user will consume the entire demand immediately in time slot t, as delaying the consumption will not increase the utility or decrease the payment.

When the prices are time and location dependent, the user may choose to schedule the traffic to later time slots (and hence at possibly different locations based on his mobility) to maximize his payoff. Let us denote the traffic that a type a user shifts from time t and location l to time t' and location l' to be $Z_{t,l}^a(t',l')$. All possible traffic shift decisions form a vector:

$$\boldsymbol{Z}_{t,l}^a = \{Z_{t,l}^a(t,l)\} \cup \{Z_{t,l}^a(t',l'), \forall t' \in \mathcal{T}_t, \forall l' \in \mathcal{L}\}. \quad (4)$$

Recall that $\mathcal{T}_t = \{t+1, t+2, \ldots, T_t\}$. Basically, if the user decides to transmit in the current time slot t, then the location can only be l (which is already known). We denote this amount of traffic as $Z_{t,l}^a(t,l)$. If the user chooses to transmit in one of the future T_t time slots, then the possible locations will be determined by the mobility pattern. Note that if the mobility pattern suggests that the user will never go to a position l' at time slot t', then we can simply let $Z_{t,l}^a(t',l') = 0$.

Furthermore, we assume that a user's demand does not change through scheduling (in the expected sense), i.e.,

$$x_0^a(t,l) = Z_{t,l}^a(t,l) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') \cdot Z_{t,l}^a(t',l'). \quad (5)$$

Here $\beta_{t,l}^a(t',l')$ is local mobility profile defined in (2).

Next we characterize how the user will calculate the utility based on the scheduled traffic. We assume that a type a user has a utility function $u_a(\cdot)$, which is an increasing and concave function of the amount of data consumed in a single time slot. Hence with the scheduled traffic, the user's utility will be:

$$u_a(Z_{t,l}^a(t,l)) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') \delta_a^{t'-t} u_a(Z_{t,l}^a(t',l')).$$

Such a utility calculation captures the decrease of utility due to delay through the parameter δ. The user's (expected) usage-based payment with the scheduled traffic will be:

$$p(t,l) Z_{t,l}^a(t,l) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') p(t',l') Z_{t,l}^a(t',l').$$

The user's objective is to maximize the his payoff (utility minus payment) by choosing the best traffic scheduling decision, i.e., solving the following optimization problem:

Problem 1: A Type a User's Traffic Scheduling Problem

$$\max \ \Big[u_a(Z_{t,l}^a(t,l)) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') \delta_a^{t'-t} u_a(Z_{t,l}^a(t',l'))\Big]$$
$$- \Big[p(t,l) Z_{t,l}^a(t,l) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') p(t',l') Z_{t,l}^a(t',l')\Big]$$

Subject to: Constraint (5),

$$\boldsymbol{Z}_{t,l}^a \geq \boldsymbol{0}. \quad (6)$$

Variables: $\boldsymbol{Z}_{t,l}^a$ defined in (4).

Since the utility function $u_a(\cdot)$ is concave, Problem 1 is a convex optimization problem. Therefore, the KKT conditions of Problem 1 is sufficient and necessary for its global optimality:

$$p(t,l) - u_a'(Z_{t,l}^a(t,l)) + \lambda_a(t,l) \geq 0, \quad (7)$$
$$\beta_{t,l}^a(t',l')\big[p(t',l') - \delta_a^{t'-t} u_a'(Z_{t,l}^a(t',l')) + \lambda_a(t,l)\big] \geq 0, \quad (8)$$
$$\boldsymbol{Z}_{t,l}^a \ \geq \ \boldsymbol{0}, \quad (9)$$
$$x_0^a(t,l) = Z_{t,l}^a(t,l) + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \beta_{t,l}^a(t',l') Z_{t,l}^a(t',l'), \quad (10)$$
$$Z_{t,l}^a(t,l) \cdot \big[p(t,l) - u_a'(Z_{t,l}^a(t,l)) + \lambda_a(t,l)\big] = 0, \quad (11)$$
$$Z_{t,l}^a(t',l') \cdot \Big[\beta_{t,l}^a(t',l')\big[p(t',l') - \delta_a^{t'-t} u_a'(Z_{t,l}^a(t',l'))\big] + \beta_{t,l}^a(t',l') \lambda_a(t,l)\Big] = 0. \quad (12)$$

For the above KKT conditions, $t \in \mathcal{T}_0, l \in \mathcal{L}, t' \in \mathcal{T}_t, l' \in \mathcal{L}, a \in \mathcal{A}$. Here $\lambda_a(t,l)$ is the multiplier corresponding to the equality constraint (5) in Problem 1, and the inequality constraint (9) is component-wise.

For the rest of the paper, we assume that the utility function is linear, i.e., $u_a(x) = \rho_a x$. Since the objective of Problem 1 is not strictly concave, the global optimal solution of Problem 1 might not be unique, which implies that a user may have more than one optimal scheduling decisions. To overcome this technical difficulty, we further assume that the operator can guide the user to choose one particular solution that the operator prefers. This does not affect the user's maximum achievable payoff, but will make the analysis easier later on.

IV. Cellular Operator's Decision in Stage I

In Stage I, the operator needs to optimize the time and location dependent prices to minimize his cost, considering the impact on the users' scheduling decisions in Stage II.

A. The Cellular Operator's Cost

We will consider two types of cost for the network operator: the cost of provisioning demand exceeding capacity, and the price discounts paid to the users to incentivize the traffic shift.

When the data traffic exceeds the network capacity at a particular time slot and location, the operator will incur significant additional cost to accommodate the extra traffic. Such a cost can be in several forms: (i) some of the traffic may not be delivered immediately, hence the users will experience a degraded QoS due to an excessive delay, which in turn may lead to user churn and reduce the operator's revenue in the long run; (ii) the operator may need to obtain additional network resources at an extra cost, such as offloading to WiFi networks belonging to a different operator, or temporally leasing spectrum from other cellular operators [15]. When the total scheduled user demand at time slot t and location l is $x(t,l)$, the cost of satisfying additional demand exceeding a capacity C is [10]:

$$f(x(t,l)) = \gamma \cdot \max\{x(t,l) - C, 0\}.$$

Here γ is the unit cost for serving an additional unit of traffic beyond the capacity.

When the cellular operator incentivizes users to shift traffic to less crowded time slots and locations through offering a price discount, the operator also experiences a loss of revenue. This can be viewed as another type of cost. Let us consider a benchmark flat-rate usage-based pricing p_0 which is time and location independent following most operators' practice today. In order to reduce the network congestion, we assume that the cellular operator can only provide time and location dependent discounts (i.e., $p(t,l) \le p_0$). Since our time and location aware pricing is given in (3), then the discount at time t and location l is $p_0 - p(t,l) \ge 0$. The constraint of providing discounts ensures that the new pricing scheme can only reduce the cost of the users. Since not providing any discounts is also a feasible pricing decision of the cellular operator, the operator will also get a cost no larger than the benchmark case with optimized discounts. Hence the "discount-only" pricing scheme can lead to a win-win situation for both the operator and users.

When the price for time slot t and location l is $p(t,l)$, the loss of revenue of serving the scheduled traffic from type a users is:

$$(p_0 - p(t,l)) \cdot x^a(t,l).$$

Here $x^a(t,l)$ is the usage for the type a user after scheduling.

B. The Cellular Operator's Price Optimization

The cellular operator's goal is to minimize his expected total cost across all time slots, locations, and user types, considering the global mobility pattern $\boldsymbol{\alpha}$ defined in (1).

Let us denote the optimal solutions of Problem 1 as vector $\boldsymbol{Z}_{t,l}^{a*}$, which is a function of the prices $\boldsymbol{p}$. Hence the usage after scheduling can be denoted as $\boldsymbol{x}^{a*}$. We assume that the user's scheduling is only for the "initial" traffic under the time and location independent pricing scheme, so the decision for time slot t does not influence his decision in the next time slot. After scheduling, the optimal usage at time slot t and location l for the type a user can be expressed as:

$$x^{a*}(t,l) = Z_{t,l}^{a*}(t,l) + \sum_{t''=\max\{t-T+1,1\}}^{t-1} \sum_{l''=1}^{L} \beta_{t'',l''}^{a}(t,l) \cdot Z_{t'',l''}^{a*}(t,l). \quad (13)$$

Problem 2: The Operator's Price Optimization Problem

$$\min \sum_{t=1}^{T_0} \sum_{l=1}^{L} \alpha_{t,l} \cdot \left[f(\sum_{a=1}^{A} x^{a*}(t,l)) - p(t,l) \cdot \sum_{a=1}^{A} x^{a*}(t,l) \right]$$

Subject to: Constraint (13),

$$0 \le p(t,l) \le p_0, \ t \in \mathcal{T}_0, l \in \mathcal{L}. \quad (14)$$

Variables: $p(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}$.

Notice that we remove the term $p_0 \cdot \sum_{t,l,a} x^a(t,l) = p_0 \cdot \sum_{t,l,a} x_0^a(t,l)$ from the objective of Problem 2, as it is a constant. Only prices are the decision variables in Problem 2, as $\boldsymbol{Z}_{t,l}^{a*}$ and $\boldsymbol{x}^{a*}$ are all functions of the prices (by solving Problem 1). However, in general it is not possible to obtain the closed-form of $\boldsymbol{Z}_{t,l}^{a*}$ and $\boldsymbol{x}^{a*}$ in terms of prices, which leads to great difficulty in solving Problem 2 directly. This motivates us to look at an equivalent bilevel problem formulation.

V. SOLVING THE BILEVEL OPTIMIZATION PROBLEM

The two-stage decision problems (Problem 1 and Problem 2) can be equivalently reformulated as a single bilevel optimization problem. In a bilevel optimization problem, a lower-level problem is embedded in an upper-level optimization problem. In this paper, the cellular operator's pricing problem (Problem 2) is the upper-level one, and users' scheduling problem (Problem 1) is the lower-level one.

When the lower-level problem is convex, its optimal solution can be characterized by the necessary and sufficient KKT conditions, which can be embedded into the high-level problem and lead to the bilevel optimization formulation [16]. By substituting the KKT conditions (7)–(12) into the operator's pricing Problem 2, we obtain the bilevel problem:

Problem 3: Bilevel Pricing and Scheduling Problem

$$\min \sum_{t=1}^{T_0} \sum_{l=1}^{L} \alpha_{t,l} \left[f(\sum_{a=1}^{A} x^{a}(t,l)) - p(t,l) \sum_{a=1}^{A} x^{a}(t,l) \right]$$

Subject to: $(7) - (12), (13), (14)$.

Variables: $p(t,l), \lambda_a(t,l), Z_{t,l}^a, x^a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}$.

Since Problem 3 is a quadratic program with linear complementarity constraints, it is *NP-hard*, and there is no polynomial time algorithms that can find the global optimal solution [17]. This motivates us to develop an easily implementable low complexity algorithm.

Since the complementarity constraints (11) and (12) are not easy to handle, we propose to penalize them to the objective function. This transforms Problem 3 to Problem 4.

Problem 4: Penalty-based Problem

$$\min \sum_{t=1}^{T_0} \sum_{l=1}^{L} \alpha_{t,l} \left[f(\sum_{a=1}^{A} x^{a}(t,l)) - p(t,l) \sum_{a=1}^{A} x^{a}(t,l) \right] + \tau \cdot \sum_{a=1}^{A} \sum_{t=1}^{T_0} \sum_{l=1}^{L} \left[\phi_{t,l,t,l}^{a} + \sum_{t'=t+1}^{T_t} \sum_{l'=1}^{L} \phi_{t,l,t',l'}^{a} \right]$$

Subject to: $(7) - (10), (13), (14)$.

Variables: $p(t,l), \lambda_a(t,l), Z_{t,l}^a, x^a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}$.

Here $\phi^a_{t,l,t,l}$ and $\phi^a_{t,l,t',l'}$ correspond to the complementarity constraints:

$$\phi^a_{t,l,t,l} = Z^a_{t,l}(t,l) \cdot \left[p(t,l) - u'_a(Z^a_{t,l}(t,l)) + \lambda_a(t,l)\right],$$
$$\phi^a_{t,l,t',l'} = Z^a_{t,l}(t',l') \cdot \Big[\beta_{t,l}(t',l')\big(p(t',l') - \delta_a^{t'-t} u'_a(Z^a_{t,l}(t',l'))\big) + \lambda_a(t,l)\beta_{t,l}(t',l')\Big].$$

Problem 4 is equivalent to Problem 3 as long as the penalty parameter τ is sufficiently large. This is because in Problem 4 we are trying to minimize two terms, one is the original objective function of total cost, and the other is the linear complementary equations. Intuitively, when the penalty parameter is sufficient large, the latter one will dominate the former one, and we are actually minimize the latter one with higher priority and the former one with a lower priority.

We have the following theorem, whose proof can be found in [18].

Theorem 1: There exists a $\tau_0 > 0$, such that Problem 4 is equivalent to Problem 3 for any $\tau \geq \tau_0$.

The threshold value τ_0 is unknown in practice, hence we need to choose the value of τ through an trial and error process. We can start by an initial estimation of τ, and solve Problem 4 (using the algorithm discussed below) until convergence. Then we can check whether the complementarity constraints (11) and (12) are satisfied at the solution. If yes, then we are done. Otherwise, we need to increase τ and then solve Problem 4 again.

Although Problem 4 still is nonconvex due to the non-convex objective function, all of its constraints are convex. This motivates us to use the block coordinate descent (BCD) algorithm to solve Problem 4 [20]. The key idea of the BCD algorithm is to partition variables in Problem 4 into two blocks: $\{p(t,l), \lambda_a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}\}$ and $\left\{Z^a_{t,l}, x^a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}\right\}$. When we fix the variables in one block, Problem 4 becomes a linear programming problem of variables in the other block, and hence can be solved efficiently (to its optimality). Then we iteratively solve the variables in two blocks until the algorithm converges.

Algorithm 1: BCD Algorithm for Problem 4

Input: $\{x^a_0(t,l), \alpha(t,l), \beta^a_{t,l}(t',l'), \delta_a, \rho_a : \forall t \in \mathcal{T}_0, t' \in \mathcal{T}_t, l, l' \in \mathcal{L}, a \in \mathcal{A}\}, C, T, p_0, \gamma, \tau, \varepsilon_0$.
Output: $\{p(t,l), x^a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}\}$
Initialization: $\varepsilon = \infty$;
while $\varepsilon > \varepsilon_0$ **do**
  Solve Problem 4 in terms of variables $\{p(t,l), \lambda_a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}\}$, assuming other variables are fixed;
  Solve Problem 4 in terms of variables $\left\{Z^a_{t,l}, x^a(t,l) : \forall t \in \mathcal{T}_0, l \in \mathcal{L}, a \in \mathcal{A}\right\}$, assuming other variables are fixed;
  ε is the relative tolerance of the new and old $Z^a_{t,l}$.
end

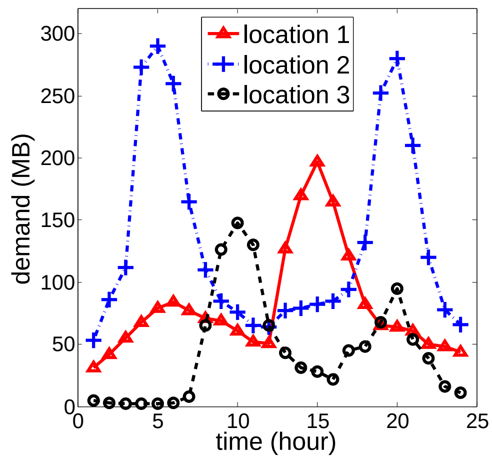

Figure 3: "Initial" Traffic of 24 Time Slots.

The complexity for the first linear programming problems is $O((T_0L(aT_0L + a + 2))^{1.5}(T_0L(a+1))^2)$, and the complexity for the second linear programming problem is $O((2aT_0^2L^2 + T_0L)^{1.5}(aT_0^2L^2)^2)$. The complexity of the entire BCD algorithm is still an open problem [19].

The proposed BCD algorithm can be easily implemented, since we can use a mature linear programming solver to solve each step. The algorithm is guaranteed to converge to a KKT solution, which is in general the best we can do for the general NP-hard problems.

Theorem 2: The sequence generated by Algorithm 1 globally converges to a KKT point of Problem 4.

Proof: Detailed proof can be found in [20]. ■

Intuitively, the variables in the constraints are decoupled after the partition, and we can solve each subproblem to the global optimality. This leads to convergence of the BCD algorithm.

VI. Simulation Results

In this section, we verify the effectiveness of the proposed pricing scheme.[2]

Figure 3 shows the initial traffic pattern under the time and location independent pricing benchmark. We can see that the peak hours (when the demand exceeds capacity) are location-dependent. We assume that the network capacity $C = 100$, the operator's pricing span $T_0 = 24$, users' scheduling interval $T = 12$, users' utility parameter $\rho = 1.1$, and the time and location independent price benchmark $p_0 = 1$.

We simulate two different cases. In the first case, we have $\gamma = 30$ and $\delta = 0.95$, and our proposed time and location aware pricing leads to a shifted traffic pattern as shown in Fig. 4. In the second case, we have $\gamma = 10$ and $\delta = 0.7$, and the corresponding shifted traffic pattern is shown in Fig. 5. Figure 6 shows the operator's optimal time and location aware prices for both cases, in which the upper one is for Setting I, and the lower one is for Setting II.

[2]Numerical results show that the BCD algorithm can achieve the global optimality for small-scale problems by using the branch and bound algorithm as the benchmark.

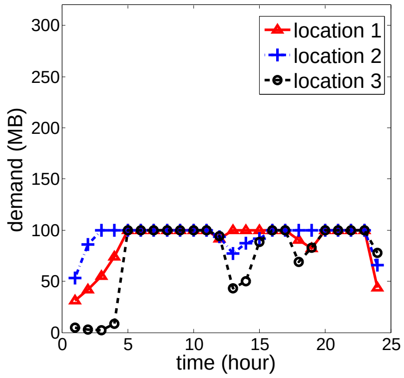

Figure 4: The Usage under Setting I.

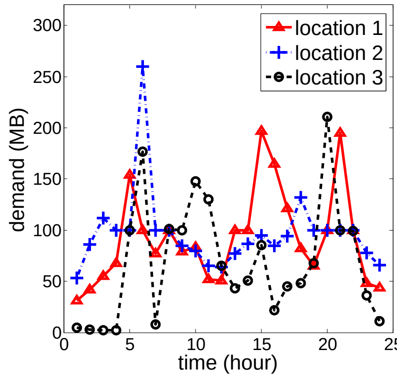

Figure 5: The Usage under Setting II.

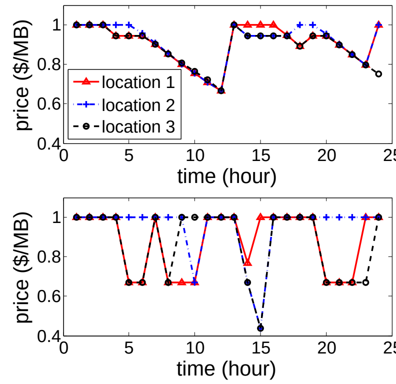

Figure 6: Prices under Two Settings.

We can get some useful observations from the simulation results. First, how much the traffic can be smoothed heavily depends on the system parameters. In the first case where $\gamma = 30$ and $\delta = 0.95$, the traffic variance is decreased by 85.28%. In the second case where $\gamma = 10$ and $\delta = 0.7$, the traffic variance is decreased by 48.39%. This means that a larger cost parameter γ and a larger delay tolerance δ makes the traffic smoothing easier to achieve. Second, our proposed pricing scheme leads to a win-win situation for both the operator and users. The cellular operator can decrease its total cost (consisting of the cost of demand exceeding capacity and the loss of revenue due to discounts) by 98.70% in the first case and 74.76% in the second case. Mobile users can increase their payoffs by 106.10% in the first case and 124.17% in the second case.

VII. Conclusion

In this paper, we provide the first study that rigorously optimize the time and location aware pricing scheme for wireless mobile data networks. We model the interaction between the cellular operator and users as a two-stage decision process, and formulate the overall pricing and scheduling problem as a bilevel optimization problem. We propose an easily implementable algorithm that involves the penalty method and the block coordinate descent algorithm to solve the bilevel problem. We verify the effectiveness of our model (e.g., up to 98.70% total cost decrease and 106.10% payoff increase). Our next step plan is to conduct large scale comprehensive simulation studies of the algorithm performance based on realistic mobile data usage traces, and create mobile apps to further understand and help users making automated traffic scheduling decisions.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017", Feb. 2012.

[2] Bloomberg, "AT&T to Pay 1.93 Billion for Qualcomm Mobile Spectrum", 2010, [Online]. Available: http://www.bloomberg.com/news/2010-12-20/at-t-agrees-to-acquire-wireless-licenses-from-qualcomm-for-1-93-billion.html

[3] Rysavy Research, "Mobile Broadband Capacity Constraints and the Need for Optimization", 2010, [Online]. Available: http://www.rysavy.com/Articles/2010_02_Rysavy_Mobile_Broadband_Capacity_Constraints.pdf

[4] Alcatel Lucent, "White Space and Mobile Comminucations", *Strategic White Paper*.

[5] J.M. Dyaberi, et al, "Managing Cellular Congestion Using Incentives", *IEEE Communications Magazine*, vol. 50, no. 11, pp. 100–107, Nov. 2012.

[6] "AT&T Mobile Share Value Plans", [Online]. Available: http://www.att.com/shop/wireless/data-plans.html#fbid=1tEwGqOhPi2

[7] "Content for All: The Potential for LTE Broadcast/eMBMS", *White Paper*, Jan. 2013, [Online]. Available: http://www.qualcomm.com/sites/ default/files/document/files/igr_qlabs_lte_broadcast_white_paper_final1.pdf

[8] U. Paul, A.P. Subramanian, M.M. Buddhikot, and S.R. Das, "Understanding Traffic Dynamics in Cellular Data Networks", *Proc. of IEEE INFOCOM*, 2011.

[9] E. Cantor, et al, "The Mother of Invention: Network Operators in the Poor World are Cutting Costs and Increasing Access in Innovative Ways", *The Economist*, Sept. 2009.

[10] C. Joe-Wong, S. Ha, and M. Chiang, "Time-Dependent Broadband Pricing: Feasibility and Benefits", *Proc. of IEEE ICDCS*, 2011.

[11] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Pricing for Demand Shaping and Proactive Download in Smart Data Networks", *Proc. of IEEE INFOCOM*, 2013.

[12] L. Law, "Game Theoretic Analysis of Congested Wireless Networks", *Ph.D. thesis, The Chinese University of Hong Kong*, 2012.

[13] J. Ghosh, M.J. Beal, H.Q. Ngo, and C. Qiao, "On Profiling Mobility and Predicting Locations of Campus-wide Wireless Network Users", *State University of New York at Buffalo, Tech. Rep.*, CSE-2005-27 (2005).

[14] G. Loewenstein, and D. Prelec, "Anomalies in Intertemporal Choice: Evidence", *Choice over time*, 1992.

[15] Techweekeurop, "EC Calls For Spectrum Sharing To Ease Mobile Data Strain". [Online]. Available: http://www.techweekeurope.co.uk/news/ec-spectrum-sharing-white-spaces-91152

[16] B.Colson, P.Marcotte, and G.Savard, "An Overview of Bilevel Optimization", *Annals of Operations Research*, vol. 153, no. 1, pp. 235–256, Sep. 2007.

[17] P.M. Pardalos and S.A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard", *Journal of Global Optimization1*, vol. 1, no. 1, pp. 15–22, Mar. 1991.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, U.S.A. Springer, 2006.

[19] A. Saha and A. Tewari, "On the Nonasymptotic Convergence of Cyclic Coordinate Descent Methods", *SIAM Journal on Optimization*, vol. 23, no. 1, pp. 576–601, Nov. 2012.

[20] L. Grippo and M. Sciandrone, "On the Convergence of the Block Nonlinear Gauss–Seidel Method under Convex Constraints", *Operations Research Letters*, vol. 26, no. 3, pp. 127-136, Apr. 2000.