Time-Dependent Pricing for Multimedia Data Traffic: Analysis, Systems, & Trials

Soumya Sen*, Carlee Joe-Wong†, Sangtae Ha ‡, and Mung Chiang §

* University of Minnesota, ssen@umn.edu ‡ University of Colorado Boulder, sangtae.ha@colorado.edu † Carnegie Mellon University, cjoewong@andrew.cmu.edu § Purdue University, chiangm@purdue.edu

Abstract

The explosive growth of multimedia data traffic in wired and wireless networks have led Internet service providers (ISPs) to use penalty mechanisms like throttling, capping, overage fees to manage network congestion. But such measures are harmful to the Internet ecosystem. Therefore, we use ideas from economics to create incentive-based, as opposed to penalty-based, solutions for data plans. In particular, we explore time-dependent pricing (TDP) - a form of dynamic pricing that manages congestion by offering time-varying discounts to incentivize users to shift some data traffic temporally. To realize TDP data plan in practice, we provide (i) an optimization model to compute time-dependent prices, (ii) a system implementation for deployment in operational networks, and (iii) experiments with two cellular networks for demonstrating feasibility. Our results show that users respond to such pricing plans by using higher volume of traffic in lower-priced (off-peak) periods and benefit from a lower $/GB fee, while the ISPs benefit from a higher revenue due to increase in off-peak usage and lower peak-to-average traffic ratio in their network. This suggests that such a pricing solution can incentivize users to modify their usage behavior and enable better revenue management in multimedia-rich networks.

I. Introduction

The growth of smart mobile devices (e.g., smartphones, tablets), bandwidth-hungry applications (e.g., ultra-high definition video streaming), cloud-based services (e.g., file backup, file sharing), multimedia-rich web content, machine-to-machine traffic from the Internet of Things, etc, have led to increased congestion in both wired and wireless networks. For example, the global mobile data traffic is projected to increase by nearly 10-fold between 2014 and 2019 [1], a compound annual growth rate of 57%. In the US, the estimated wireless spectrum crunch will

be about 500 MHz by 2019 and this gap will continue to widen over time. These developments pose a serious threat to the economic viability of the Internet. Even 5G networks will need to cope with the greater bandwidth requirements of emerging applications like telemetry, virtual and augmented reality, and autonomous vehicles/IoT. Internet Service Providers (ISPs) contend that these developments require changes in pricing to help manage congestion on cellular networks. Regulatory bodies like the FCC in the US have also noted that they *"...recognize the importance and value of business-model experimentation"* [2]. But several recent attempts by ISPs to monetize their network (e.g., paid prioritization, zero-rating, multiple tiers) have been criticized by net neutrality advocates and the Open Internet Rules [3] due to price discrimination.

Similarly, researchers have pointed out that simple usage-based pricing is unlikely to help these ISPs alleviate peak congestion on their network. For instance, Internet pioneer, Vinton Cerf, advocated that *"Network management also should be narrowly tailored, with bandwidth constraints aimed essentially at times of actual congestion"* [4]. Odlyzko et al., [5] argued that prices will need a temporal component in order to shift demand, i.e., they should vary over different times of the day, as in time-dependent pricing (TDP). TDP exploits users' time elasticity of demand for different types of multimedia traffic, incentivizing them to shift their usage from "peak" to "valley" times and improving network resource utilization. TDP is particularly suitable for ISPs because unlike unpopular penalty mechanisms, here it is the consumers who maintain control over the amount that they pay by explicitly choosing when to use data, given the prices.

TDP for mobile data is a form of variable (dynamic) pricing, which has received considerable attention in the networking literature [6]. But although the basic principles of TDP have been studied from a theoretical perspective, challenges remain in operationalizing it as a functional system in networks to study how real users actually respond to time-varying prices in mobile data plans. Previous studies have shown that pricing schemes must be designed carefully in order to elicit appropriate user responses. For example, in a trial of dynamic congestion pricing for computer-telephony (i.e., TDP for voice calls), price increases in real time did not induce users to terminate their calls early [7]. Consequently, TDP for mobile data must overcome the users' dislike of price uncertainty in real time pricing. In this work, we consider *dynamic day-ahead time-dependent usage-based pricing*[1]. With dynamic day-ahead TDP, the ISP computes hourly

[1]For brevity, we will often refer to it simply as TDP (time-dependent pricing) and the offered prices will include a discount on the pre-existing baseline usage-based fee (e.g., $10/GB).

prices based on predicted usage over the next day, posting the effective prices for all users 24 hours in advance and thus maintaining a sliding one-day window of announced prices. This ability to adjust price points each day allows the ISP to monitor changes in users' demand in response to the past prices and adapt future prices accordingly. Users, in turn, can use a pricing app on their mobile device to view and react to the current and future prices in each hour (e.g., by preemptively decreasing their current usage in anticipation of future low prices).

To realize this idea, we address three key challenges – (i) how to compute an optimized set of future prices to offer? (ii) how to design a system to operationalize this pricing plan for an ISP? and (iii) how will users understand and respond to this kind of dynamic pricing? First, we introduce an economic model and algorithms for computing time-dependent prices. This model accounts for the ISP's tradeoff between the costs of providing off-peak hour incentives and overshooting the capacity during the peak hours. We also model users' usage volumes and willingness to delay usage for different types of traffic, aggregated across all users, so as to maintain user privacy and algorithmic scalability. Second, we implemented this pricing algorithm in a system prototype, including server-side modules for measurements and price computation, and a user-side app for displaying the time-dependent prices. Third, we evaluate our system in two field trials. The first trial with AT&T wireless users demonstrates the efficacy of optimized dynamic day-ahead TDP. A second trial, conducted with an Alaskan mobile network operator, is a randomized field trial of non-optimized TDP that reinforces our findings about user response.

In addressing these three research dimensions, we make a few key contributions to the literature. Although our field studies are limited in scale due to financial and regulatory concerns, the prototype and its trials provide valuable validation of the system and insights into user behavior. In both of these trials we find quantitatively as well as qualitatively that dynamic day-ahead TDP changes user behavior: users not only decrease their usage in response to higher prices, but also preemptively increase their usage in anticipation of future high prices. We also introduce a system prototype that can serve as a model for future pricing experiments in operational networks. Thus, our work bridges the gap between the theory and practice of bringing together multimedia and economics by designing an end-to-end pricing solution for the Internet. Given the growing demand for multimedia data traffic and spectrum crunch, the realization and demonstration of these ideas are particularly timely and of interest to service providers, regulators, and consumers, as well as for application in related areas, such as the energy market [8], smart grids, IoT, and transportation networks.

The paper is organized as follows: in Section II, we discuss the related literature on network usage and pricing. We then introduce a demand-side model in Section III, and use it to optimize the dynamic TDP in Section IV. Section V discusses the system architecture and client-side user interfaces for operationalizing these pricing plans. We report on an optimized pricing trial with AT&T's users in Section VI and on a randomized TDP field experiment with customers of an Alaskan ISP in Section VII. Section VIII concludes the paper.

II. Related Literature

A. Dynamic Pricing

Our work contributes to the literature on the economics of broadband pricing, in particular, dynamic pricing. Sen et. al., [6] provides a detailed survey of various static and dynamic pricing schemes in the existing literature. Unlike static pricing, dynamic pricing promotes more efficient allocation of limited resources by designing incentives that account for the time-varying nature of congestion in the network. Many recent works [9]–[11] have therefore explored analytical and simulation-based models to investigate the impact of incentivizing off-peak capacity usage. Others have explored TDP's ability to regulate peak demand. For example, [12] introduced an analytical model to compare the consumer surplus and resource utilization under flat-rate and usage-based TDP, while [13] simulated the benefits of combining both spatial and temporal traffic patterns in time-dependent data plans. Similarly, [14] used simulation to demonstrate pre-scheduling delayed flows to improve resource utilization. While these works demonstrate the importance of dynamic TDP as a potential mechanism for regulating network congestion, they share two key limitations. First, these theoretical models have not been implemented in operational cellular networks; hence, they do not address system design or implementation challenges (e.g., the need for modifications to the core and edge network infrastructure). Second, these pricing schemes have not been tested with real users; hence, they do not capture the behavioral factors that influence the efficacy of any such incentive schemes in the real world.

Dynamic day-ahead pricing can also be applied to other contexts, such as smart grids [15], to improve utilization in the electricity market. While the basic intuitions are similar, the energy market has some key differences from a mobile network setting. First, energy providers have a very different supply-side model because, unlike bandwidth, energy is produced from different sources (e.g., coal, water, nuclear) that have different production costs and availability constraints for the resource. Second, on the demand-side mobile data usage is typically more bursty and

shiftable than energy needs of electrical devices. Moreover, much of the work on dynamic pricing in smart grids has also been based on simulations than actual field experiments.

B. Field Experiments in Network Pricing

While various analytical models of pricing have been proposed in the literature, field experiments to validate them have been very few due to the operational complexity in deploying new pricing plans. An early field experiment in Berkeley found that users have a psychological preference for flat prices compared to usage- or QoS-based ones [16]. However, the study did not test the effects of 'dynamic' prices on usage behavior. Another experiment on time-of-day pricing for computer-telephony services (i.e., voice calls) showed that static time-of-day pricing encouraged users to shift 30% of their voice calls from peak to off-peak hours, but that real-time dynamic congestion pricing did not induce users to terminate their calls earlier [7]. The result suggests that users have a psychological preference for having some level of certainty about the future prices on offer. Our dynamic "day-ahead" TDP plan provides that certainty by announcing the prices 24 hours in advance. Given the vastly different characteristics of landline voice and multimedia-rich data traffic, however, results from real-time pricing of voice calls cannot be used to interpret the effectiveness of day-ahead TDP in mobile data plans.

The need to conduct field experiments on dynamic time-dependent pricing is further motivated by recent studies in the literature, which show that users' reactions to usage- or QoS-based pricing in wired data networks depend on the interfaces with which users can monitor their usage behavior [17]–[19]. These studies in HCI have demonstrated how well-designed interfaces can help users adjust their demand according to the prices offered, thus simplifying complex pricing schemes. Therefore, in this work we use rigorous design principles to develop our pricing system and its related user-facing components.

Our work extends earlier smart data pricing research [20]–[22] in three key aspects. First, it introduces an updated theoretical framework to offer dynamic day-ahead time-dependent prices by developing an optimization model to compute prices based on both demand and supply side factors, including the possibility of additional demand driven by discounts. Secondly, it operationalizes the framework as a system prototype with user-side mobile apps for both iOS and Android. Third, it validates this system with two wireless ISPs - from the east and west coast of the US - whose users are shown to respond positively to dynamic TDP data plans.

III. Computing Time-Dependent Prices

We first discuss in Section III-A some practical considerations that a price computation framework should satisfy in order to be deployable in a real system. We then present a demand-side model in Section III-B and formulate the price optimization in Section III-C.

A. Model Requirements

The analytical model used for price computation by such a TDP system should work based only on the observable aggregate usage data. That is, it should avoid techniques such as deep-packet inspection (DPI) to monitor and measure data at the granularity of each individual user's usage and application types, thus protecting user privacy and maintaining scalability. In most networks, thousands of user sessions may pass through the same network bottleneck (e.g., middle-mile, base station) in any hour, making it impractical to infer utility function parameters for individual users. Moreover, the user utility in a given hour will also depend on the full range of prices over the day and the applications that they use at any given time, resulting in multi-dimensional utility functions. All these factors make user-level utility functions and subsequent parameter inference an impractical modeling approach.

Instead of modeling user-level utility functions, we estimate price- and delay-sensitivity across different traffic classes based only on the aggregate data demand. We use "waiting functions" to probabilistically model the amount of data that a user is willing to shift from one time period to another, given the set of time-dependent prices over the day. The parameters of each waiting function capture the tradeoff between a user's willingness to defer usage of some virtual traffic class and the price discounts offered. We then use nonlinear curve-fitting between the observed and expected aggregate demand to estimate the waiting function parameters.

B. Network Usage Model

To realize TDP, the ISP has to choose the incentives offered so as to maximize its profit, i.e., its revenue with TDP less the cost of overshooting network capacity. We set up this optimization by deriving an expression for X_s, the usage volume in each time period s with TDP, which will then be used to derive expressions for the ISP's revenues and costs in Section III-C.

We can calculate X_s as a function of the baseline usage at time s before TDP, plus the change in usage volume with TDP. In computing this change in usage from the pre-TDP baseline, we have to account for two possible effects: time-shifting of usage from more to less expensive

time periods, and any additional increase in usage driven by the discounts themselves. As the users' baseline data usage follows some daily patterns[2], we assume that overall usage behavior is cyclic with periodicity T (e.g, $T = 24$ for daily periodicity [23]); we relax this assumption in Section IV. Thus, we consider T time periods in each day, indexed by $s = 1, 2, \ldots, T$, each of which is associated with a time-dependent price p_s. We also normalize the price units so that the time-independent base price (i.e., which the ISP charges without TDP) is 1 (e.g., this can correspond to usage-based fees of \$10/GB). We further define the time-dependent discounts as $d(s) = 1 - p_s$, i.e., discounts off of the base price. We constrain $d(s) \geq 0$, ensuring that the ISP never offers prices above the base price. This constraint fits with our goal of helping ISPs move away from offering usage penalties to an incentive-based model for influencing changes in user behavior. While this may limit ISP profits, it is psychologically reassuring for users as they are guaranteed that the base price is the maximum they will pay at any time.

Let $V(s)$ denote the baseline usage volume at time s prior to introducing TDP. We first formulate the additional increase in usage that is driven solely by the discount offered, independent of the usage time-shifting. The exact increase cannot be directly measured, as the observed usage also includes the effect of time-shifting. Thus, we introduce a parameter α_s that models the degree to which usage volume at time s increases with the discount offered. In Section IV-B, we will discuss how the value of α_s can be estimated from measurable quantities. For any value of α_s, this term should be zero when no discount is offered so that only positive discounts lead to an increase in usage. We use a power-law[3] functional form

$$V(s)\left((1 + d(s))^{\alpha_s} - 1\right) \tag{1}$$

to represent the increase in usage at time s due to the discount at time s.

We now consider the amount of usage shifted from one time period to another. Users are heterogeneous in their willingness to shift data usage to a different time for different types of applications. However, since for privacy reasons we do not have application-specific usage data, we account for user and application heterogeneity by logically grouping the usage from different applications across the user base into a total of B virtual traffic classes, indexed by $b = 1, 2, \ldots, B$. These classes will be parameterized based on the "shiftability" of the traffic, i.e.,

[2]In some locations, weekend and weekday traffic patterns can be very different. We can thus solve for separate prices on weekends and weekdays; we omit this step here for clarity.

[3]The power-law form is consistent with the user demand functions assumed in economics literature [24].

the users of all traffic sessions belonging to a particular class are assumed to exhibit the same willingness to shift this usage in exchange for a given lower price. For example, these classes may correspond to different types of multimedia applications, such as ultra-high definition video, cloud backups, video conferencing, web browsing, etc. Modeling in terms of traffic classes also allows us to retain computational scalability in our model: instead of accounting for the behavior of each individual user, we simply estimate the behavior of each traffic class.

Users' willingness to shift their usage depends not only on the amount of time that the traffic of a given class is shifted by, but also on the amount of money that can be saved by such shifting. We therefore define the probability that users who are generating session traffic of class b will defer their traffic from time s to time t as

$$w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right), \tag{2}$$

which we call as a "waiting function." This function captures the tradeoff that users face between receiving a monetary reward and delaying their usage, as discussed in the behavioral psychology literature by [25]. The notation $(d(t)-d(s))$ is the difference in the price (or discount) per unit volume between periods t and s, and $|t-s|_T$ denotes the amount of time[4] that users wait if they defer to a future time, t. The parameter $\beta(b)$ characterizes the delay sensitivity for class b traffic. For instance, in our later examples we consider waiting functions to have the form $w_{\beta(b)}\left(\delta,\tau\right) = \mu_{\beta(b)}\max(\delta,0)\left(\tau+1\right)^{-\beta(b)}$, where $\mu_{\beta(b)} > 0$ is a normalization parameter to ensure that this probability function value lies in $(0,1)$. This functional form ensures that the probability of waiting typically decreases rapidly with time to wait τ; a smaller $\beta(b)$ parameter indicates more willingness to delay traffic in class b. Similar functional forms have been used in economic theory to model time-discounting of user utilities [26], [27].

We now use these waiting functions to find the amount of usage shifted into and out of a given period s. Let each traffic class b correspond to a fraction $\rho_b(s)$ of the total traffic volume, so that the expected amount of usage in traffic class b shifted from time s to time t is $V(s)\rho_b(s)w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)$. Thus, the change in usage for time s due to time-shifting of usage is $\sum_{b=1}^{B}\sum_{t=s-T+1}^{s-1}V\left(t\right)\rho_b\left(t\right)w_{\beta(b)}\left(d(s)-d(t),|s-t|_T\right) - \sum_{b=1}^{B}\sum_{t=s+1}^{s+T}V\left(s\right)\rho_b(s)$

[4]If $t>s$, then $|t-s|_T = t-s$, but if $t<s$, we assume that users defer from time s to time t on the next day. The time between these two periods is then $t+T-s$. For modeling purposes, we assume that since users know what data they want to consume at the current time, they can delay it to future time periods to get the price discounts. But they do not shift their usage from a future periods into the current period because their future usage amounts are unknown and hard to shift preemptively.

$w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)$, where the first sum represents the volume of usage shifted into time s from other time periods and the second sum the usage shifted out of time s to other periods. Combining this expression with the increase in usage (1), the total usage[5] at time s with TDP is

$$\begin{aligned} X(s) =& V(s)\left(1+d(s)\right)^{\alpha_s} + \sum_{b=1}^{B}\sum_{t=s-T+1}^{s-1} V(t)\,\rho_b(t)\,w_{\beta(b)}\left(d(s)-d(t),|s-t|_T\right) \\ & - \sum_{b=1}^{B}\sum_{t=s+1}^{s+T} V(s)\,\rho_b(s)\,w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right). \end{aligned} \tag{3}$$

C. Profit Maximization

Maximizing the ISP's profit with TDP is equivalent to optimizing the change in profit relative to that without TDP. We find the ISP's change in revenue and operational cost as a function of the discounts offered and then show that maximizing this change in profit with respect to the time-dependent discounts is a convex optimization problem, which can be solved efficiently.

We first calculate the change in revenue ΔR compared to the revenue under time-independent pricing, which is $\left(\sum_{s=1}^{T} V(s)\right)$ (with $p(s)$ is normalized to 1). There are two sources of this change in revenue: lost revenue due to the discounts offered, and a gain in revenue due to additional demand generated by the discounts over the baseline usage. To calculate the loss in revenue, we first note that if no shifting occurs, the ISP loses revenue $V(s)d(s)$ in each period s due to the discount offered. We now adjust this term to account for traffic shifted from period s to other times t. The difference in the discount per unit usage between time s to time t is $d(t)-d(s)$, and the amount of traffic shifted to time t is $\sum_{b=1}^{B} V(s)\rho_b(s)w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)$. Thus, the total loss of revenue from traffic in each time s may be expressed as $V(s)d(s) + \sum_{t\neq s}\left(d(t)-d(s)\right)\sum_{b=1}^{B} V(s)\rho_b(s)w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)$. The gain in revenue due to the traffic increase can be calculated by multiplying the price $(p_s = 1-d(s))$ with the increase in usage at time s. From (3), the revenue gain is therefore $\left(1-d(s)\right)V(s)\left(\left(1+d(s)\right)^{\alpha_s}-1\right)$, and the total change in revenue is

$$\begin{aligned} \Delta R\left(d(1),\ldots,d(T)\right) = \sum_{s=1}^{T}\Bigg[& V(s)\big(1+\left(d(s)-1\right)\left(1+d(s)\right)^{\alpha_s}\big) \\ & + \sum_{t\neq s}\left(d(t)-d(s)\right)\sum_{b=1}^{B} V(s)\rho_b(s)w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)\Bigg]. \end{aligned} \tag{4}$$

[5]Additional formulation details are provided in Appendix B.

We now find the ISP's operational cost under time-dependent pricing. We suppose that, at each time s, the ISP's network has effective capacity C_s, which is the available capacity (i.e., total capacity minus any background traffic and other traffic not charged according to the time-dependent prices. If the usage at any time exceeds the network capacity, then the ISP must accommodate this extra usage, incurring a cost $g_s\left(\max\left(X(s)-C_s,0\right)\right)$, where the network usage $X(s)$ is given by (3). For instance, the function g can include the cost of increased consumer complaints or churn due to congestion.

As the network usage $X(s)$ increases, we expect the ISP cost $g_s\left(\max\left(X(s)-C_s,0\right)\right)$ to also increase. For simplicity, we take g_s to have a constant marginal cost γ_s; the ISP's total cost due to congestion can then be written as

$$G\left(d(1),\ldots,d(T)\right)=\sum_{s=1}^{T}\gamma_s\max\left(X(s)-C_s,0\right) \tag{5}$$

The change in capacity cost relative to that without TDP is $G-\sum_s\gamma_s\max\left(V(s)-C_s,0\right)$. Since the cost without TDP is a discount-independent constant, we can omit it from the ISP's optimization problem. Maximizing the ISP's total change in profit is then equivalent to minimizing the sum of the revenue loss ΔR in (4) and the total cost G in (5):

$$\begin{aligned}\min_{d(s)}\quad & \sum_{s=1}^{T}\Bigg[\gamma_s\max\left(X(s)-C_s,0\right)+V(s)\left(1+\left(d(s)-1\right)\left(1+d(s)\right)^{\alpha_s}\right)\\ & \qquad +\sum_{t\neq s}\left(d(t)-d(s)\right)\sum_{b=1}^{B}V(s)\rho_b(s)w_{\beta(b)}\left(d(t)-d(s),|t-s|_T\right)\Bigg] \qquad (6)\\ \text{s.t. } & d(s)\geq 0,\ s=1,2,\ldots,T. \qquad (7)\end{aligned}$$

As discussed above, we restrict the discount variables to be nonnegative to ensure that users will not pay more for data usage under TDP than they would have before TDP.

Proposition 1. *The optimization problem (6–7) is convex if $\gamma_s\leq 2/(1-\alpha_s)$ for all times s.*

The proof of Proposition 1 is provided in Appendix A.

Note that as $\alpha_s\to 1$, i.e., the additional demand increases linearly with the discount, the marginal cost of extra capacity, γ_s, can be arbitrarily large. Typically, such additional demand generated will be relatively small ($\alpha\in[0,1)$), requiring that $\gamma_s\leq 2$ in order for (6–7) to be convex. Since we normalize the maximum base price to 1, we effectively constrain the marginal cost of exceeding capacity to be no more than twice the marginal price charged to users. As

the marginal cost is generally less than the marginal price to ensure a positive ISP profit, this condition will hold true and the optimization will be convex for all practical purposes.

As it is formulated as a convex optimization problem, (6–7) can be solved rapidly to calculate the optimized discounts over the day using standard optimization techniques. The convex structure of this formulation also ensures that our pricing algorithms are scalable to multiple traffic classes and time periods. We elaborate on these pricing algorithms in the next section.

IV. Pricing Algorithms

In Sections IV-A and IV-B, we provide price-computation and estimation algorithms needed to employ the model developed in Section III. We also demonstrate these algorithms' efficacy and scalability in Section IV-C using numerical evaluation with some real data.

A. Dynamic Day-Ahead Pricing

We use the framework of dynamic day-ahead TDP to relax the assumption that users' behavior is perfectly cyclical from day to day. Over time, users' attitudes towards their data usage can change, e.g., as new apps become popular. Such changes in user behavior are reflected in our model through updates to the parameter values of the waiting functions. Thus, while the discounts that minimize (6) can serve as a starting point for determining time-dependent prices, these discounts must also evolve over time. In other words, the prices and waiting function parameters will be updated dynamically over some pre-defined time window. For example, in day-ahead pricing, the ISP will compute each price point a day in advance using updated parameters in its user behavior model. We introduce the following procedure for determining the offered prices:

Step 1. The ISP conducts a pre-rollout pilot study to gather usage data and estimate the initial user behavior parameters α_s, $\rho_b(s)$, and $\beta(b)$ in (3). We suggest methods for this estimation in Sec. IV-B.

Step 2. Given these parameters, the ISP calculates the optimal discounts for times $s = 1, 2, \ldots, T$ as in (6).

Step 3. Users view these prices on their mobile devices and adjust their usage accordingly.

Step 4. After each period s, the ISP solves for the optimal discount $d(s+T)$ using (6), given the previously computed discounts $d(s+1), d(s+2), \ldots, d(s+T-1)$.

Step 5. The ISP re-estimates the parameters $\beta(b)$, $\rho_s(b)$, and α_s, periodically (e.g., once a week). This re-estimation should only be performed once a sufficient number of data points have been gathered to avoid overfitting the data.

Step 6. At longer intervals, e.g., once a month, the ISP solves for the baseline time-independent pricing usage $V(s)$. This step should be performed less often than the parameter estimation, since we assume the baseline usage volume will remain relatively stable on a daily or weekly timescale.

By allowing for periodic re-estimation of the user behavior parameters and baseline usage, this procedure allows ISPs to optimally solve for the time-dependent prices while adjusting to changes in application traffic proportions, demand surges, and other user behavior changes.

B. Estimating User Behavior

We now consider the user behavior estimation in step 1 of Section IV-A's pricing algorithm. We must estimate three types of parameters: the fractions of traffic corresponding to each traffic class, $\rho_b(s)$; the waiting function parameters $\beta(b)$; and the time-varying discount-driven usage increase parameters, α_s. To choose the optimized parameter values, we employ nonlinear curve-fitting algorithms to compute the parameter values for which (3) holds best, i.e., to solve

$$\min_{\rho_b(s),\beta(b),\alpha_s} \sum_t \left(Y(t) - X\left(t, \rho_b(s), \beta(b), \alpha_s\right)\right)^2, \tag{8}$$

where the sum includes all times for which we have usage and price data, $Y(t)$ denotes the observed usage at time s, and $X\left(t, \rho_b(s), \beta(b), \alpha_s\right)$ denotes the right-hand-side of (3) at time t as a function of the parameter values.[6] Since this function is nonlinear in the parameter values, we cannot efficiently solve (8) for optimality. A variety of curve-fitting methods exist for such estimation problems [28], which can used to solve (8) without changing the formulation of our model. We use the commonly employed Levenberg-Marquardt algorithm [29], a hybrid of the gradient descent and Gauss-Newton algorithms. This algorithm has the additional advantage of being iterative: as user behavior changes over time, we can quickly re-estimate parameter values using our previous estimates as "warm start" initial points. These behavior changes can reflect both short-term fluctuations, e.g., users' patience levels varying from day to day, as well as longer-term changes, e.g., when users begin to use data in different ways.

The system can be recalibrated over time by re-estimating the baseline usage with time-independent prices, $V(s)$, as in Step 6 of Section IV-A's pricing algorithm. An obvious, albeit impractical, way to do so is to actually offer time-independent prices to some users periodically and observe their usage levels. But this method risks confusing users accustomed to time-dependent pricing, and furthermore will erode ISPs' gains from offering TDP in the first place. As an alternative, we note that if accurate parameter values are known, (3) is linear in the baseline

[6]We note that the usage $X\left(t, \rho_b(s), \beta(b), \alpha_s\right)$ at time t is a function of the parameter values $\rho_b(s)$, $\beta(b)$, and α_s at all times s, not just $\rho_b(t)$ and α_t.

usage values $V(s)$. Thus, by fixing the parameters to their most current estimated values and taking $X(s)$ as a function of the baseline usage volumes $V(s)$, (3) becomes a system of linear equations. Solving for the new $V(s)$ is then a quadratic minimization problem

$$\min_{V(s)} \sum_{t} \left(Y(t) - X\left(t, V\left(s\right)\right)\right)^2, \tag{9}$$

which may be easily solved with standard algorithms.

The estimations in (8) and (9) require specifying a functional form for the waiting functions $w_{\beta(b)}$. While any functional form can be used with this formulation, we assume the following conditions to make it reasonably consistent with expected user behavior. First, if no change in discounts is offered, users do not shift any of their traffic, as they have no incentive to do so: $w_{\beta(b)}\left(0, \tau\right) = 0$ when $\tau > 0$. Second, for a fixed time to wait τ, users' willingness to wait is increasing in the change in discounts δ: users are more likely to defer their traffic if they receive a larger monetary reward for doing so. Finally, we assume that for a fixed change in discounts δ, $w_{\beta(b)}$ is decreasing in τ: users are less likely to wait for a longer period of time. We also suppose that users will not shift their traffic for more than one day: $w_{\beta(b)}\left(\delta, \tau\right) = 0$ if $\tau > T$.

C. Numerical Evaluation

We first demonstrate the accuracy of Section IV-B's user behavior estimation using a network usage dataset from a wireless ISP. We then use these parameters to perform a sample day-ahead price calculation using the algorithms of Section IV-A. The runtimes of both algorithms are well within operational requirements, even with many traffic classes and time periods in a day.

Behavior estimation efficacy: In this example, we consider waiting functions of the form

$$w_{\beta(b)}\left(\delta, \tau\right) = \frac{\mu_\beta \max(\delta, 0)}{\left(\tau + 1\right)^{\beta(b)}}, \tag{10}$$

where μ_β is an appropriate normalizing constant[7] to map $w_{\beta(b)}$ in $(0, 1)$. These functions satisfy the requirements that $w_{\beta(b)}\left(0, \tau\right) = 0$ and that $w_{\beta(b)}$ is increasing in δ and decreasing in τ. Figure 1a shows the cumulative normalized root-mean-square error in usage estimation over three days when we estimate the waiting function parameters based on the preceding 30 days of data from 7 users who received random time-dependent prices for 33 days (Section VII). The

[7]We can define $\mu_{\beta(b)} = \sum_{\tau=1}^{T} (\tau+1)^{-\beta(b)}$, i.e., the sum of all possible waiting function values given a maximum difference in usage fees of 1.

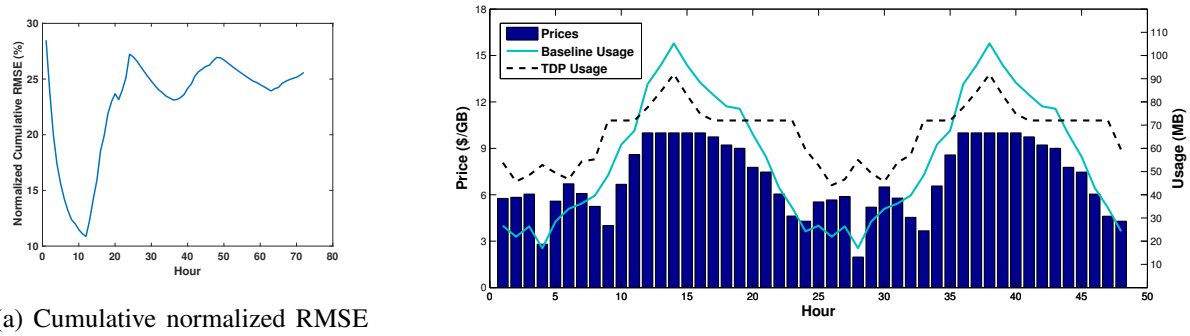
error is consistently between 10% and 30%, indicating that our estimates are accurate even on the relatively noisy usage data generated by our 7 users.

To show that our estimation is accurate in settings beyond this user trial, we consider a model of $T = 24$ periods and $B = 3$ traffic classes. We take the first, second, and third traffic classes to have respectively low, medium, and high delay sensitivities: $\beta(1) = 0.5$, $\beta(2) = 1$, and $\beta(3) = 3$. We use data from 27 customers of an Alaskan ISP (cf. Section VII) to calculate the time-independent traffic volumes $V(s)$ and proportions of traffic $\rho_b(s)$ corresponding to each traffic class for each period. Traffic class 1 corresponds to downloads, traffic class 2 to video streaming, and traffic class 3 to social networking and emails. The α_s parameters are chosen from a uniform distribution between 0 and 1.

We simulate two weeks of data usage patterns with random time-dependent discounts chosen from a uniform distribution between 0 and 1 and $T = 24$. We then calculate the resulting traffic volumes $X(s)$ according to (3) with the parameters and usage volumes above, and mimic real user behavior by adding a random variable of $x(s)$, drawn from a normal distribution with mean 0 and variance $X(s)/2$, to each $X(s)$. The mean percentage difference $|x(s)/X(s) - 1|$ between the perturbed and ideal usage was 8.77%. After estimating the parameters according to (8), we find that the calculated traffic volumes $Y(s)$ with the estimated parameters are very close to the actual time-dependent traffic volumes $X(s)$, with a mean percentage error of 2.16%. The estimated α_s values are within 0.1 of the true values 90% of the time. The estimated $\beta(b)$ values are $(\beta(1), \beta(2), \beta(3)) = (0.513, 1.113, 2.951)$, which are quite close to the actual values of $(0.5, 1, 3)$ that were used in simulating the data usage pattern under TDP.

Price calculation: We next calculate the optimized day-ahead prices over two days of traffic, using the user behavior parameters above. We consider the ISP cost parameters for capacity of a bottleneck pipe and the cost of overshooting it to be $C_s = 72$MB and $\gamma_s = \$50$/GB for all times s. Figure 1b shows the optimal prices and the resulting usage in each hour, compared to the pre-TDP usage baseline. The simulation demonstrates that the algorithm to compute TDP offers higher prices in peak periods to reduce network congestion and lower prices in valley periods to incentivize shifts in demand, thereby creating a flatter demand profile for better resource utilization, i.e., with decreases in the peak usage and increases in off-peak usage.

We further evaluate the effect of the length of the pricing period: while hourly prices are intuitively easy for users to understand, changing the prices less frequently (e.g., peak-load pricing, which uses two peak/off-peak periods) may also reduce network congestion. We simulate



(a) Cumulative normalized RMSE over 3 days of usage estimation.

(b) Two days of optimal prices and simulated usage.

Fig. 1: (a) Usage estimation and (b) TDP simulation on data from an Alaska user trial (Section VII). Our estimation has relatively small errors (under 30%), and TDP reduces the peak usage and fills up the valley periods.

offering 3, 12, and 24 prices per day (corresponding to 8, 2, and 1 hour length periods) and find that our pricing plan with hourly prices reduces traffic peaks most significantly; it leads to a peak-to-average traffic ratio of 1.3625, compared to 1.4455 for 2-hour prices, and 1.5648 for 8-hour prices. More details on these simulations are given in Appendix D.

Algorithm scalability: We implement our behavior estimation and optimal price calculation algorithms in Matlab and Python and evaluate their runtimes as we increase the optimization complexity, i.e., the number of traffic classes and periods in a day. The price computation of each future period should finish within the duration of the current period, and the user behavior estimation, which is run daily, should finish within 24 hours. Theoretical complexity bounds for these algorithms depend on the specific optimization method used. For instance, using the subgradient method [30] to find the optimal prices has complexity $O(\epsilon^{-2}T^2)$, where ϵ is the permitted error, for finding T optimal prices (step 2 of Section IV-A's algorithm); subsequent day-ahead price calculations (step 4 of Section IV-A's algorithm) have complexity $O(\epsilon^{-2}T)$ as only one price needs to be optimized. Using the Levenberg-Marquardt algorithm for the behavior estimation (steps 5 and 6 of the algorithm) has complexity $O(\epsilon^{-2}T^3B^3)$ [31]. We emphasize that we ran basic implementations of our algorithms on a commodity Intel Xeon server; using faster hardware or refining the optimization algorithms used will likely decrease the runtimes significantly. More details on the complexity bounds are given in Appendix C.

First, we measure the computational overhead (the total runtime in Matlab) as we increase the number of periods from 12 to 144 (2 hour to 10 minute periods). Table Ib shows the measured

	Number of Traffic Classes		
Periods	2	4	8
12	0.21	12.99	21.52
24	3.33	47.08	75.47
48	15.99	197.22	215.42

(a) Behavior estimation (minutes).

Number of Periods	12	24	48	96	144
Behavior Estimation	12.76	200.0	959.6	1967	15040
Price Calculation	1.67	1.69	1.70	1.81	1.84

(b) Behavior estimation and price calculation (seconds).

TABLE I: Runtime of the behavior estimation and price calculation algorithms.

run-time of the parameter estimation and price calculation algorithms. Even with 144 periods, the price calculation is quite fast; the estimation algorithm performs adequately, as it runs only once a day for day-ahead TDP. We also measure the effect of adding traffic classes to the behavior estimation, as shown in Table Ia. The computation with 48 periods and 8 traffic classes still takes less than 4 hours (215.42 minutes), which is more than fast enough, as the estimation runs once a day. Our estimation uses one month of simulated data, which was generated by perturbing the usage predicted from given waiting functions by up to 50%. The reported running times were averaged across five computations with random data and starting points.

V. System Design

Implementing a new pricing mechanism in a network operator's billing system requires incorporating new functionalities such as real-time usage measurement and price computation. We develop an architectural framework to separate these functionalities between the system backend (in the network core) and the end-user devices (at the network edge), which shown in Figure 2. In the system backend, *i.e.,* on the ISP-side, our system has usage measurement, user behavior estimation, and price computation modules. For the user device, we develop a mobile app that exchanges price and usage information with the system backend. The design and evaluation of the resulting IT artifact also offers a prototype for conducting network pricing experiments.

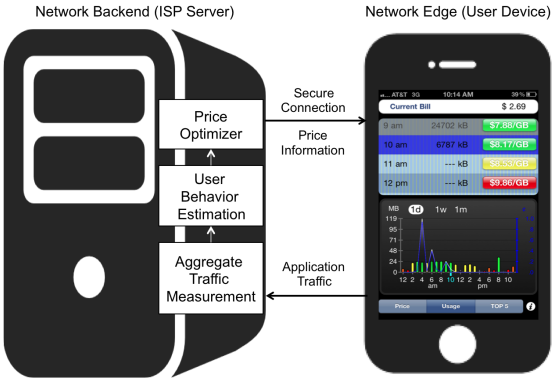


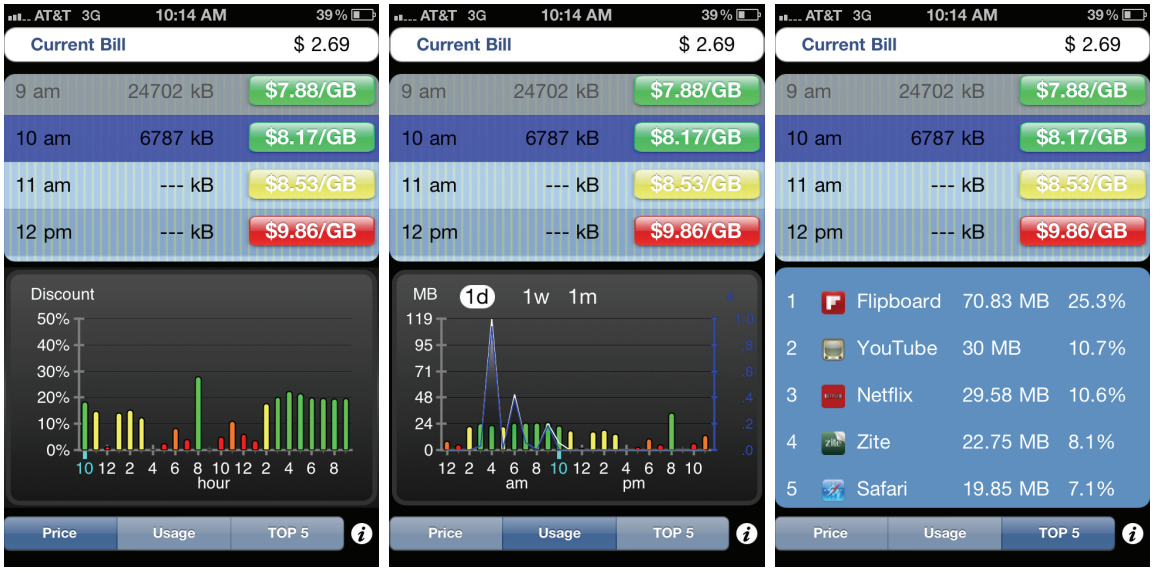
Fig. 2: Functionality separation between the network backend and edge devices.

Relatively few works have developed human-facing systems for field experiments with cellular networks. Some exceptions include the *Eden* system [32], which modifies a home router to provide users with an intuitive interface for managing their "home network experience." Similarly, the *Homework* project [33] modifies the handling of protocols and services in the home router to monitor data usage, prioritize different devices, and monitor other users' data consumption. These works, however, focus on user interaction and interface design. In contrast, our work focuses on studying the impact of economic incentives (i.e., pricing) on user behavior, requiring us to develop a new ISP- and user-side architecture that can be deployed in practice. The user interfaces of our mobile pricing application incorporated recommendations from usability research and were refined through multiple rounds of focus group studies.

A. System Architecture

Network backend system: We implemented the network backend components on a Linux system with an Intel Xeon 2.0 GHz CPU and 8GB of RAM. We provide a standard web-based API so that any device supporting web connectivity can exchange data with the ISP server. The implementation is in Python using the Django and MySQL DB.

Mobile app: We designed and implemented a mobile application that runs on end-user devices (i.e., at the network edge) for both the iOS and Android platforms. This app pulls the price information from the ISP server at regular intervals over a secure TLS/SSL connection and displays the prices for the next 24 hours to the users. The prices on offer are shown in a color-



(a) Price display. (b) Price and usage. (c) Top 5 applications.

Fig. 3: Screenshots of the TDP app on an iPhone. iPhone users can a) check the prices for next 24 hours, b) learn from price and usage history, and c) identify top 5 apps by bandwidth usage.

coded format[8], e.g., red for a high price (no discount) and green for a low price (50% discount). The app also displays the user's mobile data usage history by day, week, and month, as well as information about the top five bandwidth consuming apps. These interfaces are shown in Figure 3. The current price is displayed in a color-coded indicator in the top bar of the device's home screen so that users can view this price without having to launch the app.

In the following sections of the paper, we show how this system was deployed in two field trials. The first trial offered optimized dynamic day-ahead time-dependent pricing to a group of AT&T users using the price computation algorithms introduced in Section IV. The second trial was conducted with an Alaskan ISP for which a non-optimized TDP plan was deployed to perform exogeneous price variations to understand the degree to which individuals are actually responsive to these interventions, thus demonstrating the generalizability of the results.

[8]In order to ensure that the design does not disadvantage color-blind users, we also provide the price information as % discounts and as $/GB as secondary signals.

VI. Trial 1: Optimized Dynamic Day-Ahead Time-Dependent Pricing

In this section, we provide an overview of an optimized dynamic day-ahead TDP trial conducted with AT&T cellular users. For brevity, we will refer to the trials as TDP trials. We expect that the optimized TDP will offer lower prices in off-peak periods, i.e., those with lower usage, and higher prices at peak times. We discuss our experimental setup and data collection methods before presenting our trial results, which show that TDP can benefit both users and ISPs.

A. Trial Setup

Network setup: The technical intricacies of operational cellular networks introduce challenges for conducting pricing experiments. AT&T provided an Access Point Name (APN) that allowed us to separate the participants' 3G data traffic from all other AT&T customers' traffic[9]. After this setup, AT&T securely tunneled trial participants' 3G traffic from its 3G core network to our lab server that offered the time-dependent prices.

Software installation: The trial participants installed a custom profile on their iPads that specified a custom APN name, allowing data traffic from the iPads to reach the APN that AT&T created for the trial. The custom profile also disabled WiFi so that participants did not offload their traffic onto a supplementary network.As shown in Figure 3, the participants also installed our mobile app on their devices to view the prices offered from our server. In order to add certain features (e.g., color-coded price discount information on the home screen), we had to jailbreak all of the iPads to gain root access to these devices[10].

Money flow: We recruited 9 participating households in Princeton, NJ, using mailing lists. Like other studies with expensive methodologies (e.g., neuroIS experiments [34]), the scale of the experiment was limited by financial and regulatory constraints, as discussed in Section VIII. Our participants had diverse ages, genders, and professions. Additionally, we conducted a second trial of similar scale in Alaska (Section VII) to demonstrate generalizability of the key insights.

Each participant was leased an iPad2 with AT&T connectivity for the duration of the trial. To ensure that users' behavior during the trial reflected their real usage and spending preferences, we charged users for data access according to our time-dependent pricing. These prices were

[9]We used AT&T's 3G network because its all IP-core 4G network was not operational at the trial location at that time.

[10]In the follow up trial with users of an Alaskan operator (Section VII), we allowed users to use WiFi and tracked WiFi availability in our app. Jailbreaking their devices was not a practical option, and hence we used a simpler TDP realization.

offered one day in advance and changed in each hour (i.e., $T = 24$ in Section IV's pricing model). We paid AT&T according to its actual data plan (a flat rate of $32 for 2GB and $10/GB for additional data for each iPad), thus acting as a resale ISP offering TDP to the trial participants.

Data collected: We collected 78 days of pre-trial data and 11 days of TDP trial data at an hourly granularity. During the trial, we recorded the price offered and each user's volume of cellular data usage in each hour of the day. We did not record per-app usage data due to privacy concerns, but did display this data to users in their trial apps to help them monitor their usage.

B. Patterns in the Data: Usage and Revenue

TDP's effectiveness depends on how the offered prices impact hourly traffic volume, users' and ISPs' cost and revenues, and the network's peak-to-average traffic ratio (PAR). Ideally, TDP should lead to (i) higher traffic in lower-priced periods, (ii) a "win-win" on costs for both users and ISPs, and (iii) a lower PAR. We now assess whether our TDP system achieved these goals.

We first consider the impact of the offered prices on the hourly traffic volume: *Do users use more (less) data in the low-priced (high-priced) periods than before the trial?*

To control for differences in user behavior and hourly variations in usage over the day, we calculate the ratio of the usage during the trial and the average usage in each hour before the trial at the same hour of the day, for each user. We use logarithmic ratios to account for the long tail of the usage distribution. Figure 4 shows these ratios for different prices offered, corresponding to red, orange, and green color-coded price indicators (i.e., high, medium, and low prices) in the trial app. We see that typically during the lower-priced (green) periods, the average usage increased, while during the higher-priced (red) periods, the average usage decreased. This indicates that users responded to the prices by adjusting their usage volumes in these periods.

We next explore the impact of users' response to TDP on the $ amount paid by them per GB, and consequently, on the revenues of the ISP: *Can TDP create a "win-win" for both users and ISPs?* This "win-win" can result when a user pays a lower $/GB price for data, but increases his/her overall usage in GB in the low-priced periods, thus increasing the ISP's revenue as well. To show that this outcome occurred in our TDP trial, we first plot the amount of money that users paid for data in $/GB. Figure 5a shows that users on average paid between $7.60 and $9.20/GB over the trial period (compared to the original baseline price of $10/GB without TDP), indicating that they saved some money per GB by using more data during low-priced times. But Figure 5b shows that users also consumed a higher volume of data in GB over the month. This outcome

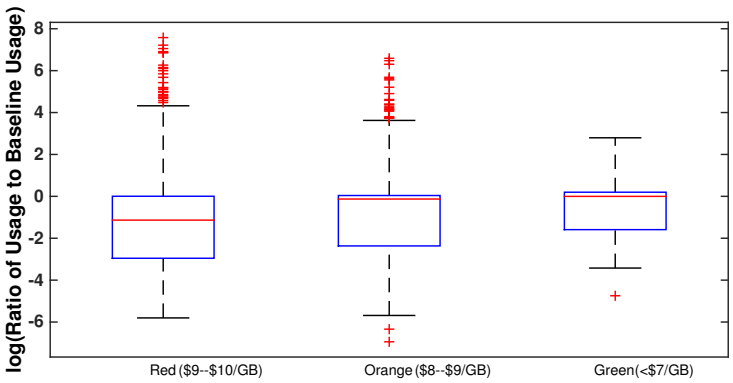
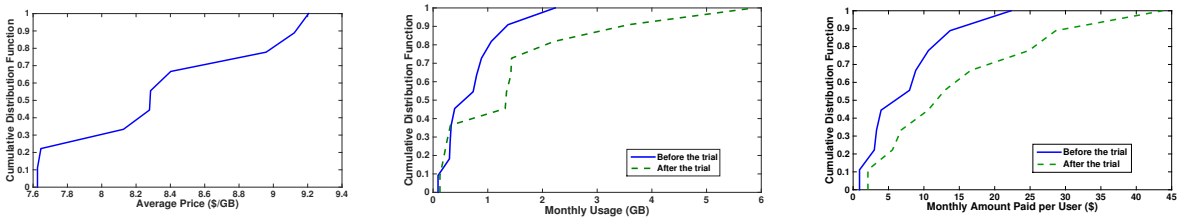


Fig. 4: Box plots of the ratio of hourly usage during the TDP trial to average hourly usage before the trial for the same user in the same hour. Mean usage is lower (higher) with higher (lower) price points.



(a) Average price per GB paid.

(b) Monthly data usage.

(c) Monthly amount paid.

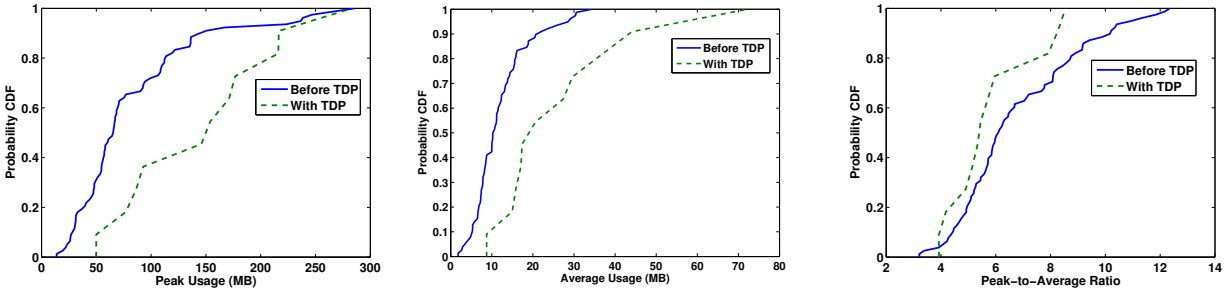
Fig. 5: With TDP, users (a) save money per GB and (b) increase overall usage, and (c) the ISP's revenue increases.

may be explained by a "sales day" effect that induces users to take advantage of the discounted low-priced periods by increasing their usage at those times (in addition to shifting traffic to these periods). As a result, the ISP's revenue also increased, as shown in Figure 5c.[11]

Given the increase in overall usage with TDP, we finally explore the trial's impact on the peak-to-average ratio of (PAR) traffic volume, i.e., *How does TDP affect the PAR?*

Figure 6 shows the distribution of daily usage statistics before and during the trial. Figure 6b shows that the average hourly usage increased during the trial, which is consistent with the increase in ISP revenue observed above. But Figure 6a shows that although the daily peak volumes during the trial may increase, the maximum peak usage does not increase. Thus, although the ISP earns a higher revenue, its capacity needs do not increase; instead, TDP leads to better

[11]Note that these results are averages: a user who remains financially conservative will enjoy a reduction in their monthly bill, while others might consume more content at a lower marginal rate, thus also helping the ISP to improve revenue and congestion.



(a) Daily peak usage. (b) Average hourly usage per day. (c) Daily peak-to-average ratios.

Fig. 6: With TDP, (a) the maximum peak usage doesn't increase although daily peaks can; (b) users increase their overall usage (in off-peak periods), leading to (c) lower PARs (i.e., flatter demand profile).

utilization of the available capacity. Furthermore, the distribution of the PARs in Figure 6c shows that they are generally smaller during compared to before the trial, indicating that usage was spread more evenly across the day. This result is consistent with greater increases in usage during lower-price times, as lower prices would be offered during less congested times. These findings are also supported by the post-trial debriefing of participants in Appendix F.

These results together indicate three key results: (i) mobile users will respond to time-dependent prices by adjusting their usage, (ii) users can benefit from a lower $/GB price, which can induce them to use more data, thus, increasing the ISP's revenue, and (iii) this additional demand will be realized mostly in low-priced periods, thus improving available resource utilization without requiring expansions to existing capacity.

VII. Trial 2: Day-ahead Time-Dependent Pricing in Alaska

Following the previous trial of the system prototype with AT&T users, we conducted a second trial with a cellular operator in Alaska to reinforce the insights on user response to time-varying prices. In this experiment, the prices offered at different times were randomized (i.e., not optimized) so as to avoid endogeneity issues in the econometric analysis. This trial also allows us to evaluate whether users with a different demographics will respond to time-varying prices in a similar manner as the AT&T users and if users shift demand in cellular networks when they have access to a secondary network, i.e., WiFi. Due to financial and regulatory constraints, we limited our field trial to 18 participants recruited by the operator from their existing user

base, who were randomly assigned to treatment and control groups. Unlike the AT&T trial with leased iPads, all participants used their personal Android smartphones throughout the trial.

A. Trial Groups and Setup

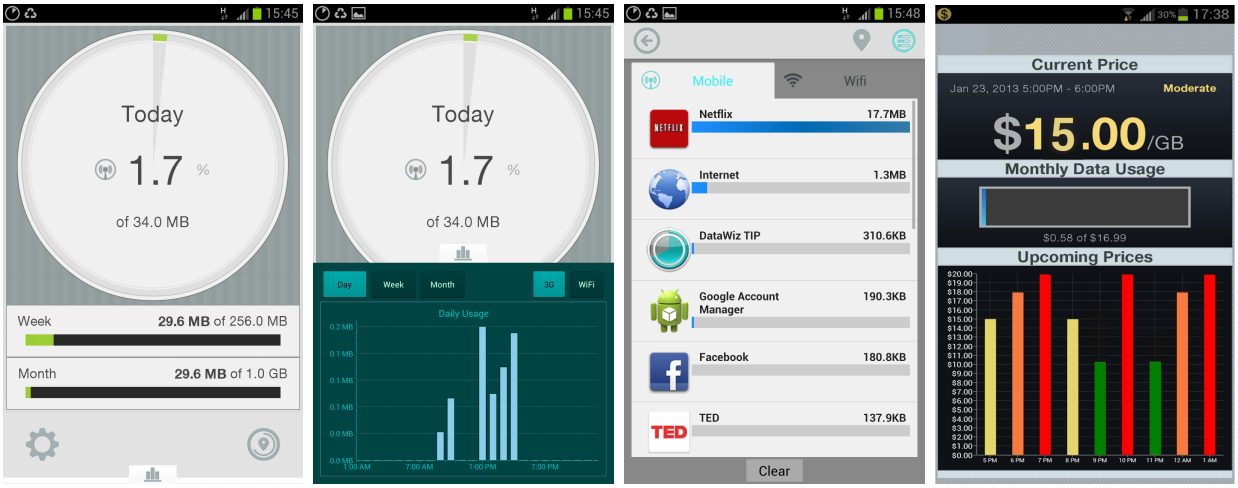
In the pre-trial phase of one year, we passively recorded users' hourly data usage volume, without changing their data plans. We then conducted a one-month pricing trial in which we randomly assigned the users into a control group and a treatment group, which received the dynamic day-ahead time-dependent prices; we refer to the treatment group as the TDP group. Thus, we have long panel data of pre-trial data over one year for these 18 users and 30 days of trial usage data, collected hourly for each user. The control group had 11 users who received a mobile app on their smartphones that monitored and displayed their data usage throughout the trial month, aiming to educate users on their behavior. Figures 7a–7c show screenshots of the app. The control group users were charged a baseline usage-based fee of $20/GB, which is the standard rate plan of the Alaskan ISP. The 7 users assigned to the TDP group[12] received a similar app, which had the same usage monitoring features as the control app and displayed hourly time-dependent prices (Figure 7d). The random price offered in each hour was chosen from four different time-dependent price points: $10/GB, $15/GB, $18/GB, and $20/GB, i.e., 0 to 50% discounts off of the $20/GB baseline price. To help users easily differentiate between the price points, each was represented in a different color (red, orange, yellow, and green for $20, $18, $15, and $10/GB respectively) on the price indicator bar in the app. At any given time, the users could view the prices for the next 24 hours, and a new price point was added every hour. All TDP users received the same set of prices, i.e., non-discriminatory at any given time.

B. Data and Descriptive Statistics

Throughout the trial phase, we recorded hourly cellular and WiFi usage volume for each user, as well as the time-dependent prices offered. Since the operator's servers recorded only cellular usage, we used the trial apps to record users' hourly WiFi usage volumes and sent these measurements to our servers[13]. We aggregate the recorded per-application usage statistics over a

[12]Originally we had 9 users in the TDP group from the randomized assignment, but 2 users were removed as their mobile apps malfunctioned. Even without these users, the control and treatment groups were statistically similar.

[13]Sending such statistics consumed negligible amounts of data and did not affect the trial results.



(a) App home screen. (b) Daily usage graphs. (c) Cellular usage by app. (d) TDP home screen.

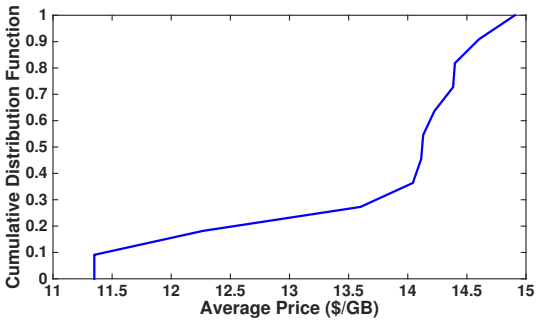
Fig. 7: Screenshots of the Android GUIs of the trial's control and TDP apps.

day for privacy preservation; hence, we do not measure the price elasticity of demand for each individual application. Both groups of trial participants, control and treatment (TDP), displayed statistically similar usage volumes and hourly usage patterns during the pre-trial period[14].

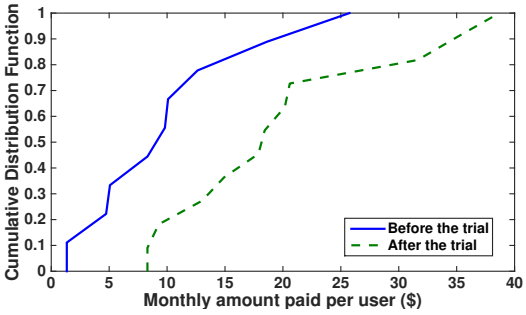
Table II summarizes the 3G data usage and WiFi availability statistics for each group of users during the trial. The TDP users have slightly higher average usage than control users, but use WiFi a little less. While the pre-trial usage volumes of the control and TDP users follow the same distribution, the distribution of the usage volume of these two groups of users during the trial should be statistically different if the TDP group users shift their usage in response to the prices. Kolmogorov-Smirnov tests find statistically significant differences between the control and TDP users' usage distributions at each of the four discount levels offered, suggesting that the time-dependent prices induced TDP users to change their usage behavior.

As in the optimized AT&T trial, we find that both users and the ISP benefit from TDP. Figure 8a shows that all users paid $< \$15$/GB, indicating that they saved money with TDP on a \$/GB basis. However, Figure 8b shows that the operator's revenue also increased because of additional demand being generated in the discounted hours, indicating a "win-win" scenario.

[14]We find that average pre-trial daily usage volume is similarly distributed among users of both groups: both samples follow a normal distribution (the Lilliefors test fails to reject the null hypothesis with significance $p > 0.1$ for both groups), and the unpaired t-test fails to reject the null hypothesis at a significance $p = 0.511$.



(a) Average price per GB paid for each TDP user.



(b) Monthly amount paid for each TDP user.

Fig. 8: With time-dependent pricing, (a) users save money per GB, and (b) the ISP's overall revenue increases.

TABLE II: Descriptive statistics: control vs. treatment.

	Control			Treatment (TDP)			
Variable	Mean	St. Dev.	N	Mean	St. Dev.	N	Total N
Hourly Trial Usage	0.547	2.618	8,712	0.652	3.577	5,544	14,256
Price Offered	–	–	–	15.001	4.598	792	792
WiFi Available*	0.39	0.488	8,712	0.266	0.442	5,544	14,256

*1 = some WiFi used; 0 = not used. N = number of data points

C. *Econometric Specification and Results*

The data collected from the experiment is studied using regression analysis – a statistical method used in econometrics to estimate the relationships among variables. Regression allows one to understand how a dependent variable (DV) changes when any one of the independent variables (IVs) or predictors is varied, while the other IVs are held constant. In the context of this experiment, the usage volumes at different times of the day serve as the dependent variable. We include several independent variables. First, data usage tends to be lower night and higher during the day [23]. To account for this periodicity, we employ an Auto-Regressive Distributed Lag (ARDL) model[15] [37]. Including a lagged usage variable also allows for temporal interdependence of usage in consecutive periods if users continue their data sessions from one period to the next. Second, the usage volume in a given hour with day-ahead TDP will depend on

[15]While a lagged dependent variable and individual fixed effects may bias the estimation of the coefficients on the order of $1/T$ [35], where T is the length of the time dimension, the length of our time dimension mitigates any such concerns [36].

both the current and future prices: users may increase their usage in anticipation of high future prices, or decrease their usage to wait for lower future prices. A generalized ARDL model [38] with lags of independent variables up to r hours in the past and price variables up to q hours in the future is therefore used. Third, users with access to WiFi would likely use WiFi rather than cellular data as WiFi data does not count towards their data plans. Lastly, users may be differently disposed to adjust their usage in response to the prices at different hours.

We estimate a log-log model to obtain percentile price elasticity and account for skewed usage distributions across users (i.e., some users have very large usage volumes as indicated by the large standard deviations in usage in Table II) [23]. Our model specification is

$$\log U_{it} = \alpha + \beta \log p_{it} + \gamma_{ih} h_{it} + \delta_{ih} h_{it} \log p_{it} + \eta w_{it} + \lambda_{ih} h_{it} w_{it} + \sum_{s=1}^{24} \nu_s \log U_{i,t-s} + \sum_{s=1}^{23} \mu_s \log p_{i,t+s} + \epsilon_{it} \tag{11}$$

where i indexes the user and t indexes the hour. User i's cellular usage volume at time t is denoted as U_{it}, with corresponding price p_{it}, allowing us to differentiate between the prices offered to the control and TDP users. WiFi access for user i at time t is denoted by a dummy variable w_{it}. The coefficients for the price variables in the current and future hours respectively are denoted by β and μ_s, h_{it} is the vector[16] of user-hour fixed effects, which account for temporal variation in user heterogeneity, and ϵ_{it} denotes user-hour specific errors.

We include terms to capture possible interactions between the user-hour dummies and the price offered at time t. We do not include an interaction term between the prices and WiFi availability since regardless of the prices offered, users will typically choose to use WiFi rather than cellular data if possible. The availability of WiFi, however, can vary over the day. Availability of WiFi at a given hour can have a multiplicative effect on user demand (e.g., WiFi availability at 3 am may not impact usage while WiFi availability at 3 pm can). Therefore we include an interaction term between WiFi availability and the user-hour fixed effects. We also include the usage volumes in the previous 24 hours of the day because of the day-ahead pricing.

[16]For each user i, the vector h_{it} consists of 24 indicator variables; when a given data point is from the hour $t = \tau, \tau \in (1, 24)$, then $h_{i\tau} = 1$ and $h_{it \neq \tau} = 0 \; \forall \; t \in (1, 24)$. Similarly, for each user i, γ_{ih} is the vector of 24 coefficients, one for each hour.

TABLE III: Regression results on hourly usage volume, (11) and (12).

Explanatory variable	(11)	(12)
Log(Current Price)	-0.879** (0.42)	-0.571* (0.354)
WiFi access	-2.238** (0.885)	-2.236** (0.885)
Log(Previous hour's usage)	0.602*** (0.0102)	0.603*** (0.0102)
Log(Next hour's price)	0.532* (0.289)	–
Observations	10386	10386
F-stat	44.7 (967,9576)	46 (944,9599)
R-square	0.791	0.79

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We report the results of an ARDL stepwise estimation with the addition of a WiFi availability variable. We also estimate (11) without the lagged prices to rigorously test their effect:

$$\log U_{it} = \alpha + \beta \log p_{it} + \gamma_{ih} h_{it} + \delta_{ih} h_{it} \log p_{it} + \eta w_{it} + \lambda_{ih} h_{it} w_{it} + \sum_{s=1}^{23} \nu_s \log U_{i,t-s} + \epsilon_{it} \quad (12)$$

Table III shows our estimation results for (11) and (12); the numbers reflect the value of the coefficient and standard deviation (in brackets) of the explanatory variables. The current price is negatively associated with usage at a statistically significant level: users responded to higher time-dependent discounts by increasing their usage in those periods and decreasing their usage in high-priced (non-discounted) periods, adjusting their behavior according to the current prices.

WiFi availability has a negative and statistically significant association with cellular usage, likely because users will prefer to use WiFi over cellular data if possible. Cellular usage is shown to have a positive dependence on the price in the next hour $(p_{i,t+1})$. This positive dependence indicates that a higher price in the future encourages users to increase their usage in the current time as they are likely to decrease their usage in those higher-priced hours. Conversely, a lower future price makes users decrease their usage in the current time period so as to use more in the future discounted hour. These effects, however, are weaker than the effect of the price at the current time (the coefficients for p_{t+1} are lower in absolute value than those for p_t): users are less likely to shift their usage to different times than they are to react to the current prices. This is likely because shifting their usage in response to prices at other times would require launching the trial app to view past and future prices instead of simply glancing at the price indicator bar on the top of the home screen that shows the current price. We also find that

usage in the previous hour is positively and statistically significantly associated with current usage, suggesting that users' sessions in the previous hour carried over into the current hour. A robustness check in Appendix E on Table III's regression with the top one percentile of usage volumes removed shows that the results remain qualitatively unchanged.

These results which show that mobile data users can and will respond to TDP in the desired manner help reinforce the insights from the deployment of our TDP system in the AT&T trial. Such a system can thus benefit users and ISPs while maintaining a privacy-preserving and non-discriminatory approach to network congestion management.

VIII. Discussions & Conclusion

Multimedia data from video services, bandwidth-heavy applications, content delivery networks (CDN), cloud, IoT have become a dominant source of traffic in wired and wireless networks. But the spectrum crunch resulting from this growth is driving many ISPs around the world to penalize demand, thus threatening the Internet ecosystem. Addressing this issue requires new ideas in multimedia economics - an interdisciplinary approach that combines economic models with systems implementation and experiments - to help network providers find better ways to improving network congestion, resource utilization, and revenue management. Therefore, in this work we examine the design and efficacy of time-dependent pricing plans - a form of dynamic pricing for multimedia data traffic with different price elasticities. We present an analytical and algorithmic framework for scalable computation of optimized prices, design and architect a system prototype, and deploy this data plan in the real world within operational ISPs.

From a business perspective, the study helps demonstrate that by designing such a system, it is feasible to successfully realize dynamic time-dependent pricing in broadband data plans instead of taking resort to unpopular penalty mechanisms. It also shows that such a pricing scheme has the potential to create a "win-win" for both users and ISPs; users respond to the price incentives by shifting some of their demand and benefit from a lower $/GB fee, while the ISPs benefit from a higher revenue due to increases in off-peak hour usage and a lower peak-to-average traffic ratio (PAR) in their network. The work is also useful from a regulatory perspective because this pricing scheme is net neutral and provides a way for sustainable growth of the Internet traffic which is essential to the continued growth of e-commerce, IoT and 5G services. More broadly, such dynamic pricing mechanisms are of interest even in the context of smart grids and for traffic management in transportation networks.

One area for further research in the industry will be to conduct field deployments of network pricing that are larger in scale. Our experiments were limited in size due to financial and operational constraints of an academic environment. But we gathered long panel data at hourly granularity over the trial period so as to have enough data for our quantitative analysis. Furthermore, we reinforced the insights on user response to dynamic TDP obtained from the evaluation of our prototype in AT&T's network with results from a second field experiment with an Alaskan network operator. The economic principles and system setup outlined in this work can provide a path for other researchers to conduct future experiments with alternative pricing mechanisms in operational networks.

In summary, this work takes a holistic approach in designing and studying the efficacy of a time-dependent pricing scheme for multimedia data traffic, involving the new models, system design and deployment in field trials. Such pricing schemes are likely to have significant implications for the Internet's long-term sustainability and accessibility to a wider user population. These ideas are therefore of interest not only to the networking research community but also to the broader e-commerce sector, policy makers, and the general public.

REFERENCES

[1] Cisco Systems, "Cisco visual networking index: Global mobile data traffic forecast update, 20142019," February 3 2015.

[2] J. Genachowski, "New Rules for an Open Internet," 2010, FCC statement.

[3] FCC, "In the Matter of Protecting and Promoting the Open Internet," 2015.

[4] V. Cerf, "What's a reasonable approach for managing broadband networks?" 2008, Google Public Policy Blog.

[5] A. Odlyzko, B. S. Arnaud, E. Stallman, and M. Weinberg, "Know Your Limits: Considering the Role of Data Caps and Usage Based Billing in Internet Access Service," *Public Knowledge Whitepaper*, 2012.

[6] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–37, 2013.

[7] J. Shih, R. Katz, and A. Joseph, "Pricing experiments for a computer-telephony-service usage allocation," in *Proc. of IEEE GLOBECOM*, vol. 4, 2001, pp. 2450–2454.

[8] S. Gottwalt, W. Ketter, C. Block, J. Collins, and C. Weinhardt, "Demand side management – a simulation of household behavior under variable prices," *Energy Policy*, vol. 39, no. 12, pp. 8163–8174, 2011.

[9] M. El-Sayed, A. Mukhopadhyay, C. Urrutia-Valds, and Z. J. Zhao, "Mobile data explosion: Monetizing the opportunity through dynamic policies and qos pipes," *Bell Labs Technical Journal*, vol. 16, no. 2, pp. 79–99, 2011.

[10] C. Chang, P. Lin, J. Zhang, and J. Jeng, "Time dependent adaptive pricing for mobile internet access," in *Proc. of IEEE INFOCOM Workshop*, 2015, pp. 540–545.

[11] M. Harishankar, N. Srinivasan, C. Joe-Wong, and P. Tague, "To accept or not to accept: The question of supplemental discount offers in mobile data plans," in *IEEE Conference on Computer Communications*, 2018, pp. 2609–2617.

[12] L. Zhang, W. Wu, and D. Wang, "Time dependent pricing in wireless data networks: Flat-rate vs. usage-based schemes," in *IEEE Conference on Computer Communications*, 2014, pp. 700–708.

[13] J. Ding, Y. Li, P. Zhang, and D. Jin, "Time dependent pricing for large-scale mobile networks of urban environment: Feasibility and adaptability," *IEEE Transactions on Services Computing*, pp. 1–14, 2018.

[14] M. Jin, S. Gao, H. Luo, J. Li, Y. Zhang, and S. K. Das, "An approach to pre-schedule traffic in time-dependent pricing systems," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 334–347, 2019.

[15] C. Joe-Wong, S. Sen, S. Ha, and M. Chiang, "Optimized day-ahead pricing for smart grids with device-specific scheduling flexibility," *IEEE Journal on Selected Areas in Communications*, vol. 30, pp. 1075–1085, 2012.

[16] H. Varian, "The Demand for Bandwidth: Evidence from the INDEX Project," in *Broadband: Should We Regulate High-Speed Internet Access*, R. W. Crandall and J. H. Alleman, Eds. AEI-Brookings, 2002, pp. 39–56.

[17] M. Chetty, R. Banks, A. Brush, J. Donner, and R. Grinter, "You're capped: Understanding the effects of bandwidth caps on broadband use in the home," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 3021–3030.

[18] M. Chetty, D. Haslem, A. Baird, U. Ofoha, B. Sumner, and R. Grinter, "Why is my Internet slow?: Making network speeds visible," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1889–1898.

[19] M. Chetty, R. Banks, R. Harper, T. Regan, A. Sellen, C. Gkantsidis, T. Karagiannis, and P. Key, "Who's hogging the bandwidth: The consequences of revealing the invisible in the home," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 659–668.

[20] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time-dependent pricing for mobile data," in *Proc. of ACM SIGCOMM 2012*. ACM, 2012, pp. 247–258.

[21] S. Sen, C. Joe-Wong, S. Ha, J. Bawa, and M. Chiang, "When the price is right: Enabling mobile time-dependent pricing," in *Proc. of ACM SIGCHI*. ACM, 2013.

[22] C. Joe-Wong, S. Ha, S. Sen, and M. Chiang, in *Proc. of Passive and Active Measurement*, 2015.

[23] A. Ghose and S. P. Han, "An empirical analysis of user content generation and usage behavior on the mobile internet," *Management Science*, vol. 57, no. 9, pp. 1671–1691, 2011.

[24] Y. Song, S. Ray, and S. Li, "Structural properties of buyback contracts for price-setting newsvendors," *Manufacturing & Service Operations Management*, vol. 10, no. 1, pp. 1–18, 2008.

[25] M. Scholten and D. Read, "The psychology of intertemporal tradeoffs," *Psych. Review*, vol. 117, no. 3, p. 925, 2010.

[26] J. Myerson and L. Green, "Discounting of delayed rewards: Models of individual choice," *Journal of the experimental analysis of behavior*, vol. 64, no. 3, pp. 263–276, 1995.

[27] T. L. McKerchar, L. Green, J. Myerson, T. S. Pickford, J. C. Hill, and S. C. Stout, "A comparison of four models of delay discounting in humans," *Behavioural Processes*, vol. 81, no. 2, pp. 256–259, 2009.

[28] D. M. Bates and D. G. Watts, *Nonlinear regression: iterative estimation and linear approximations*. Wiley, 1988.

[29] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.

[30] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge University Press, 2004.

[31] K. Ueda and N. Yamashita, "On a global complexity bound of the levenberg-marquardt method," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 443–453, 2010.

[32] J. Yang, W. K. Edwards, and D. Haslem, "Eden: Supporting home network management through interactive visual tools," *Proc. of UIST*, pp. 109–118, 2010.

[33] R. Mortier, T. Rodden, P. Tolmie, T. Lodge, R. Spencer, A. Crabtree, A. Koliousis, and J. Sventek, "Homework: Putting interaction into the infrastructure," *Proc. of UIST*, pp. 197–206, 2012.

[34] A. Dimoka, P. Pavlou, and F. Davis, "Research commentary-neurois: the potential of cognitive neuroscience for information systems research," *ISR*, vol. 22, no. 4, pp. 687–702, 2012.

[35] S. Nickell, "Biases in dynamic models with fixed effects," *Econometrica*, vol. 49, no. 6, pp. 1417–1426, 1981.

[36] W. H. Greene, *Econometric Analysis*. Prentice Hall, 2000.

[37] J. Parker, "Distributed lag models," Draft manuscript, 2014, http://academic.reed.edu/economics/parker/s14/312/tschapters/S13_Ch_3.pdf.

[38] R. C. Hill, W. E. Griffiths, and G. C. Lim, *Principles of Econometrics*. Wiley Hoboken, NJ, 2008, vol. 5.

APPENDIX A
PROOF OF PROPOSITION 1

Proof: To show that (6–7) is a convex optimization problem, it suffices to show that (6) is a convex function of the discounts $d(s)$.

We begin by writing out the cost G as

$$\gamma_s \max\left(X(s) - C_s, 0\right) = \gamma_s \max\Big(V\left(s\right)\left(1+d(s)\right)^{\alpha_s} + \sum_{b=1}^{m}\sum_{t=s-T+1}^{s-1} V\left(t\right)\rho_b\left(t\right) w_{\beta(b)}\left(d(s)-d(t), |s-t|_T\right) - \sum_{b=1}^{m}\sum_{t=s+1}^{s+T} V\left(s\right)\rho_b\left(s\right) w_{\beta(b)}\left(d(t)-d(s), |t-s|_T\right) - C_s, 0\Big) \tag{13}$$

We now note that one may treat each function g_s as a linear function of $X(s) - C_s$, with slope $\delta_s = 0$ or γ_s: we then derive conditions on the marginal cost δ_s such that the second derivative of $G + \Delta R$ is positive-definite, and show that these hold for both 0 and γ_s.

The terms of G that are piecewise-linear in the discounts do not affect the second derivative of G. Thus, we find that the second derivative $\partial^2 G/\partial d(s)^2$ of G with respect to the discount $d(s)$ in each period s is $\delta_s V(s)\alpha_s(\alpha_s - 1)\left(1+d(s)\right)^{\alpha_s - 2}$, and the mixed derivatives $\partial^2 G/\partial d(s)\partial d(t) = 0$. We now consider the first term in ΔR, i.e., $\sum_{s=1}^{T} V(s)\left(1 + \left(d(s)-1\right)\left(1+d(s)\right)^{\alpha_s}\right)$ and find that it has the second derivative $V(s)\alpha_s\left(\left(\alpha_s - 1\right)\left(d(s)-1\right)\left(1+d(s)\right)^{\alpha_s - 2} + 2\left(1+d(s)\right)^{\alpha_s - 1}\right)$; again, the mixed derivatives equal zero. We add these second derivative terms from Γ_1 and Γ_2 to obtain the expression

$$V(s)\left(\delta_s + d(s) - 1\right)\alpha_s(\alpha_s - 1)\left(1+d(s)\right)^{\alpha_s - 2} + 2V(s)\alpha_s\left(1+d(s)\right)^{\alpha_s - 1}.$$

To show that this expression is nonnegative, we note that since $d(s) \geq 0$ and $0 \leq \alpha < 1$, we require $\left(\alpha_s - 1\right)\left(\delta_s + d(s) - 1\right) \geq -2\left(1+d(s)\right)$. Noting that $\left(\alpha_s - 1\right)\left(\delta_s + d(s) - 1\right) \geq \delta_s\left(\alpha_s - 1\right)$ and that $-2\left(1+d(s)\right) \leq -2$, it suffices to show that $\delta_s\left(\alpha_s - 1\right) \geq -2$, or equivalently, that $\delta_s \leq 2/(1-\alpha_s)$. Since $0 \leq 2/(1-\alpha_s)$, we thus derive the condition that $\gamma_s \leq 2/(1-\alpha_s)$ for all times s.

We next consider the remaining terms of ΔR:

$$\sum_{s=1}^{T}\sum_{t\neq s}\left(d(t)-d(s)\right)\sum_{b=1}^{m} V(s)\rho_b(s) w_{\beta(b)}\left(d(t)-d(s), |t-s|_T\right).$$

Using the definition (10) for the w_β, we find that this expression equals

$$\sum_{s=1}^{T}\left(\sum_{t\neq s}^{n}\max\left(d(t)-d(s),0\right)^2 V(s)\sum_{b=1}^{m}\mu_{\beta(b)}\left(1+|t-s|_T\right)^{\beta(b)}\right).$$

It then suffices to show that $\max(d(s)-d(t),0)^2$ is a convex function of $d(s)$ and $d(t)$ for each period i and k. Letting $\theta \in [0,1]$ and $d(s_1)$, $d(s_2)$, $d(t_1)$, $d(t_2)$ denoting possible discounts in period s and period t, we must show that

$$\begin{aligned}&\theta \max\left(d(t_1)-d(s_1),0\right)^2+(1-\theta)\max\left(d(t_2)-d(s_2),0\right)^2\\&\geq \max\left(\theta\left(d(t_1)-d(s_1)\right)+(1-\theta)\left(d(t_2)-d(s_2)\right),0\right)^2. \qquad (14)\end{aligned}$$

If $\theta\left(d(t_1)-d(s_1)\right)+(1-\theta)\left(d(t_2)-d(s_2)\right)\leq 0$, then (14) immediately holds since the right hand side equals 0. If $d(t_1)\geq d(s_1)$ and $d(t_2)\geq d(s_2)$, then we let $x_1=d(t_1)-d(s_1)$ and $x_2=d(t_2)-d(s_2)$ to obtain

$$\theta x_1^2+(1-\theta)x_2^2=x_2^2+\theta\left(x_1^2-x_2^2\right)\geq\left(\theta x_1+(1-\theta)x_2\right)^2=\left(x_2+\theta\left(x_1-x_2\right)\right)^2.$$

Simplifying, we must show that

$$\left(x_1-x_2\right)\left(x_1+x_2-2x_2-\theta\left(x_1-x_2\right)\right)=(1-\theta)\left(x_1-x_2\right)^2\geq 0,$$

which is indeed the case. Finally, we check the case where $\theta\left(d(t_1)-d(s_1)\right)+(1-\theta)\left(d(t_2)-d(s_2)\right)>0$ and only one of $d(t_1)-d(s_1)$ or $d(t_2)-d(s_2)$ is positive; without loss of generality, we assume that $d(t_1)-d(s_1)>0$. Then (14) becomes

$$\left(\theta\left(d(t_1)-d(s_1)\right)+(1-\theta)\left(d(t_2)-d(s_2)\right)\right)^2\leq\theta^2\left(d(t_1)-d(s_1)\right)^2\leq\theta\left(d(t_1)-d(s_1)\right)^2,$$

and $\max(d(s)-d(t),0)^2$ is a convex function of $d(s)$ and $d(t)$. We have thus shown that $G+\Delta R$, as defined in (6), is convex in the discounts offered d_i. ■

APPENDIX B
FORMULATION DETAILS

In this section, we derive the total usage expressions (3) in the main text from a model of individual users' behavior. We consider a population of N users, indexed by $n=1,2,\ldots,N$, who use a total of B distinct traffic classes. Given a set of future day-ahead prices, each user n can decide whether to delay his or her consumption of class b traffic at time s, and if so, to which time it should be delayed. We define the random variable $\chi_{n,b,s}$ as the time to which users decide to delay the traffic. The probabilistic nature of $\chi_{n,b,s}$ allows us to account for fluctuations in a user's willingness to delay some traffic.

The distribution of the variable $\chi_{n,b,s}$ depends on the prices offered at future times $t > s$ and can be defined as our waiting functions (2):

$$\mathbf{P}\left[\chi_{n,b,s} = t | d(\cdot)\right] = w_{\beta(b)}\left(d(t) - d(s), |t-s|_T\right).$$

We now assume that each user has the same volume of different traffic types at each time,[17] which allows us to find the volume of traffic at time s corresponding to each traffic class b as $V(s)\rho_b(s)$. We can then find the change in usage in period s due to usage being shifted from one period to another:

$$\sum_{n=1}^{N}\sum_{b=1}^{B}\sum_{t:\chi_{n,b,t}=s} V\left(t\right)\rho_b\left(t\right) - \sum_{n=1}^{N}\sum_{b:\chi_{n,b,s}\neq s} V\left(s\right)\rho_b\left(s\right). \tag{15}$$

The increase in usage volume due to discounts offered can be given by (1) as in Section III. Combining (1) and (15), the amount of traffic in period s is then

$$X(s) = V\left(s\right)\left(1+d(s)\right)^{\alpha_s} + \sum_{n=1}^{N}\sum_{b=1}^{B}\sum_{t:\chi_{n,b,t}=s} V\left(t\right)\rho_b\left(t\right) - \sum_{n=1}^{N}\sum_{b:\chi_{n,b,s}\neq s} V\left(s\right)\rho_b\left(s\right). \tag{16}$$

We now make a final assumption that allows us to simplify (16), namely, that users' decisions to delay their usage are independent and identically distributed. In other words, each user n makes an independent decision to delay his or her traffic corresponding to parameter $\beta(b)$. Given a sufficiently large number of users, by the strong law of large numbers, the empirical distribution function of the $\chi_{n,b,s}$ for each fixed b and s approximates the actual distribution function of the $\chi_{n,b,s}$. This distribution is exactly the cumulative density function corresponding to the waiting function $w_{\beta(b)}$. Thus, the fraction of users for whom $\chi_{n,b,t} = s$ is $w_{\beta(b)}\left(d(s) - d(t), |s-t|_T\right)$. This observation allows us to rewrite (16) to obtain (3):

$$\begin{aligned}\mathbb{E}\left[X(s)\right] =& V\left(s\right)\left(1+d(s)\right)^{\alpha_s} + \sum_{b=1}^{m}\sum_{t=s-T+1}^{s-1} V\left(t\right)\rho_b\left(t\right) w_{\beta(b)}\left(d(s)-d(t), |s-t|_T\right) \\ &- \sum_{b=1}^{m}\sum_{t=s+1}^{s+T} V\left(s\right)\rho_b\left(s\right) w_{\beta(b)}\left(d(t)-d(s), |t-s|_T\right),\end{aligned} \tag{17}$$

[17] It is not difficult to generalize this assumption to consider users with independent, identically distributed usage volumes; we do not make this explicit for simplicity.

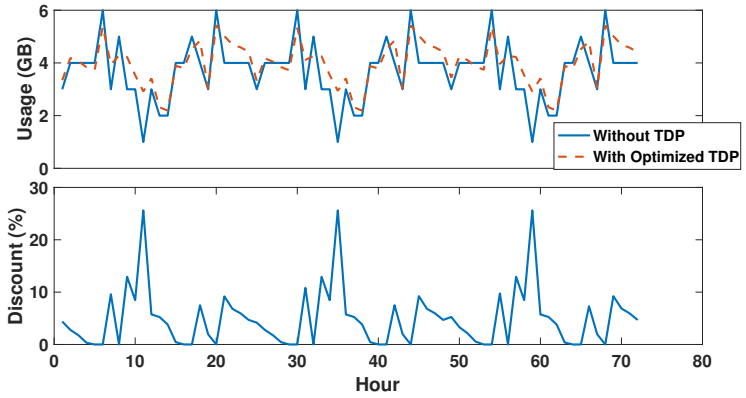
APPENDIX C
ALGORITHM COMPLEXITY

Price calculation: Our computation of the optimal time-dependent prices (steps 2 and 4 in Section IV-A's algorithm) involves solving the optimization problem (6-7). Since this problem is convex, a subgradient method will converge to the optimal solution. Computing the subgradient with respect to each optimization variable has complexity $O(T)$; since the subgradient method has complexity $O(\epsilon^{-2})$ [30], we then obtain an overall complexity of $O(T^2\epsilon{-}2)$ (step 2, solving for all T prices) or $O(T\epsilon{-}2)$ (step 4, solving for a single day-ahead price).

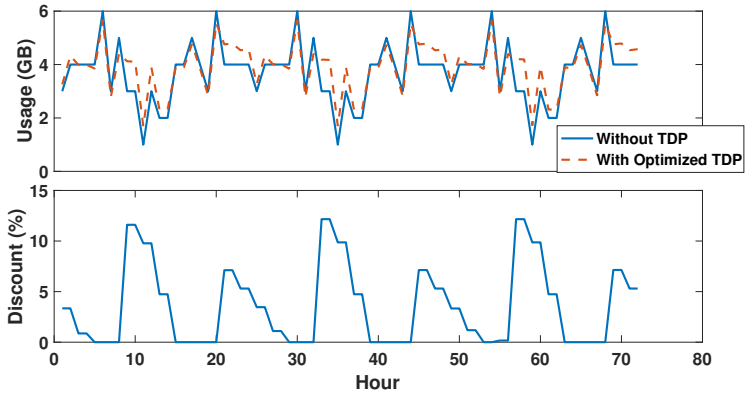
Behavior estimation: We use the Levenberg-Marquardt algorithm for our estimation, though other least-squares minimization algorithms can be used instead. This algorithm requires $O(\epsilon^{-2})$ iterations to converge to a solution with error ϵ. Thus, to find the overall complexity with respect to the number of time periods T and traffic classes B, we must find the complexity of each iteration. This is dominated by the need to compute $(J^TJ+\lambda I)^{-1}$, where J is the Jacobian of the function F, with each entry in F representing the error at time t, with respect to the optimization variables β, α, ρ; λ is a parameter chosen by the algorithm; and I is the identity matrix. The complexity of computing J is then $O(T^2B)$, since the number of optimization variables grows as $O(TB)$. Computing J^TJ given J then has complexity $O(T^3B)$, since the number of times at which we evaluate the error grows as $O(T)$. Finally, since J^TJ has dimensions of order $O(TB)$, we use the fact that matrix inversion has complexity $O(n^3)$ for an n-x-n matrix to obtain an overall complexity of $O(T^2B + T^3BT^3B^3) = O(T^3B^3)$ per iteration.

APPENDIX D
EFFECT OF VARYING PRICE PERIODS

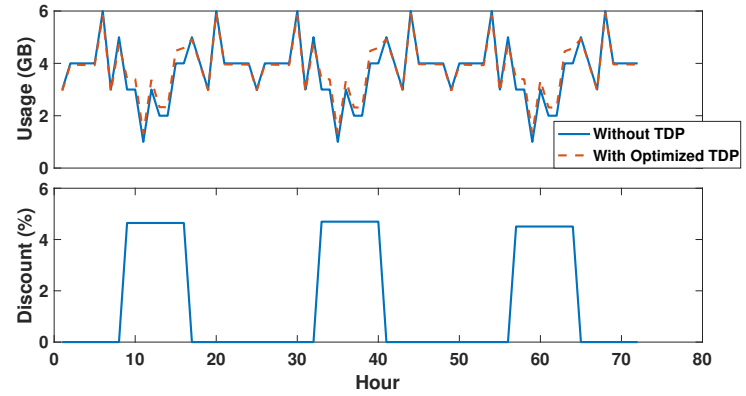
We next consider the length of the time periods in day-ahead TDP. Our choice of a 24-hour window is practical both in terms of marketing an actual data plan and for providing enough flexibility to users to plan their future usage and to ISPs to adapt their prices from day to day. However, our model can be adapted to other sizes of the sliding window of advertised prices. In this section we simulate the efficacy of some of these alternatives and find that our choice of hourly day-ahead time-dependent pricing works better. In particular, we compare alternative plans in which the offered prices last for 1, 2, and 8 hours (i.e., users receive 24, 12, and 3 prices per day respectively in these plans).



(a) Twenty-four prices per day.



(b) Twelve prices per day.



(c) Three prices per day.

Fig. 9: The plots show the usage (demand) and optimized discounts offered over three days for day-ahead prices with different numbers of prices per day, i.e., varying price advertisement window sizes. Offering hourly prices yields significantly larger peak reduction and valley filling due to the ability to adapt discounts temporally in adjacent peak and valley periods.

For the purpose of this simulation, we use the baseline hourly usage statistics from our control group (who were not offered time-dependent prices) in our Alaska trial (cf. Section VII). We scale up our baseline usage to 1000 users and assume a capacity of 4.5 GB in each hour. Next, to model how users probabilistically shift their usage in response to the prices, we assume five classes of waiting functions, corresponding to short video streaming (e.g., YouTube), streaming movies, software downloads, web browsing, and other traffic. We use the data from a survey of U.S. consumers reported in [20] to initialize the waiting function parameters β and assume linearity in the excess demand created by the price discounts (i.e., $\alpha = 1$) for these simulations.

The simulation results reported below consider three separate scenarios in which an ISP offers dynamic pricing plans in which the offered prices last for 1, 2, and 8 hours. At the beginning of the simulation, users are provided with a full day of prices, and as each price expires, a new price point is added. For instance, in the simulation with three daily prices, users are provided with the next price after 8 hours, when the first price offered expires. We run the simulation for three days for each pricing scenario, calculating the optimized prices using the framework reported in Section IV.

Figure 9 shows the resulting usage and discounts offered for each of the three pricing period durations. We observe that our pricing plan with hourly prices reduce the peaks most significantly; it leads to a peak-to-average traffic ratio of 1.3625, compared to 1.4455 for 2-hour prices, and 1.5648 for 8-hour prices. The results imply that offering prices that last for 2 or 8 hours simply does not provide enough flexibility for users to shift their data usage from one price period to another. In contrast, the price points in hourly pricing varies significantly from hour to hour, thereby exploiting the user's willingness to shift usage at a finer time granularity. Moreover, we find that the discounts offered in the two- or eight-hour pricing cases are significantly smaller than those offered for hourly pricing. This is because when prices cannot be changed at a finer granularity, the same prices prevail during both peak and valley periods, thereby moderating the effects of optimal discounts offered by the ISP.

It is also possible to offer prices more than one day in advance, e.g., for 48 hours in advance. However, doing so will reduce the ISP's flexibility to dynamically adapt its prices offered in response to changes in user demand patterns. Moreover, it is unlikely to offer additional benefits in terms of shifting user demand. With the waiting function parameters used in Figure 9, there is only about a 4% probability that users will shift their data usage by a full 24 hours; the probability that they will further delay their usage for more than one day is therefore even

smaller. Intuitively, we would expect this result, since users' daily routines and usage patterns are roughly cyclical: if they are willing to shift their usage to a another time of the day, they are likely to shift to the nearest such time, instead of waiting another day to consume this data.

APPENDIX E
ROBUSTNESS CHECKS

We now demonstrate the robustness of our results with the full econometric specification (Table III). We remove the top 1 percentile of usage volumes and show the results of regressing on (11–12) in Table IV. Compared to our original regression results in Table III, we see that there is little change in the coefficients.

TABLE IV: Regression results on hourly MTA usage volume with the top 1% of usage removed, (12) and (11).

Explanatory variable	(12)	(11)
Log(Current Price)	-0.846** (0.347)	-1.033** (0.411)
WiFi access	-2.213** (0.862)	-2.215** (0.863)
Log(Previous hour's usage)	0.593*** (0.0101)	0.593*** (0.0101)
Log(Next hour's price)	–	0.351 (0.283)
Observations	10286	10286
F-stat	47.1 (944,9499)	45.7 (967,9476)
R-square	0.796	0.796
User-hourly fixed effects	Included	Included
Price-hour interaction terms	Included	Included
WiFi-hour interaction terms	Included	Included
Lagged variables	ARDL(24,0)	ARDL (24,23)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

APPENDIX F
QUALITATIVE FEEDBACK FROM TRIAL PARTICIPANTS

Our work also contributes to ongoing research in design science by highlighting how economics and user behavior should be considered jointly to address design challenges of increas-

ingly complex socio-technical ecosystems. The qualitative observations and quotations that we present here were obtained from post-trial interviews conducted with the AT&T trial participants.

A. Attitude towards Time-Dependent Pricing

All our participants regarded the TDP data plan as viable – they would be willing to adopt it *"as long as the interface is simple to use."* But TDP's suitability for a particular person depends on the predictability of the offered prices. Some users were more price-sensitive and even adapted their online activities based on the announced prices for the day: *"I think it is a nice option to have where I can get a discount per month depending on when I use it, and I can schedule my day that way."*

Cost Savings: The participants also reported that our TDP app helped them to be more conscious in avoiding unnecessary usage at high price periods: *"Yes, and [I] was less likely to goof off and waste more time and data."* In fact, many tried to save money by limiting usage to discounted periods: *"I made a conscious effort to look for the discounts."* Moreover, they did not feel that the TDP plan required them to significantly modify their behavior: *"I go to my bank account everyday, so I would think that this would just become a natural thing."*

Sales-Day Effect in Discounted Hours: An interesting phenomenon we observed during the trial is that the high discount offers induced a 'sales-day' effect among several participants; that is, they started using more than they otherwise would have. When asked about this, one participant told us: *"[laughs] Kind of! But that also goes towards my personality of if it's on sale I must buy it!"* In fact, relative to participants' pre-TDP usage, we observed an overall increase in usage with TDP, possibly due to such a sales-day effect. This result benefits ISPs: with TDP, they can offer discounts to shift traffic from peak to off-peak periods, as well as increase demand in off-peak hours. The result is mutually beneficial, as ISPs benefit from "valley filling" and users gain by consuming more at the discounted rates in off-peak times.

These qualitative observations are consistent with the quantitative results presented in Sections VI and VII.