Joint Smart Pricing and Proactive Content Caching for Mobile Services

John Tadrous*, Atilla Eryilmaz†, and Hesham El Gamal†
*Electrical and Computer Engineering Department, Rice University, Texas, E-mail: jtadrous@rice.edu
†Electrical and Computer Engineering Department, The Ohio State University, Ohio,
E-mail: {eryilmaz.2, elgamal.2}@osu.edu

***Abstract*—In this work, we formulate and study the profit maximization problem for a wireless service provider (SP) that encounters time varying, yet partially predictable, demand characteristics. The disparate demand levels throughout the course of the day yield excessive service cost in the peak hour that substantially hurts the reaped profit. With the SP's ability to track and statistically predict future requests of its users, we propose to enable proactive caching of the peak hour demand ahead during off-peak times. Thus network traffic will be smoothed out, while end users' activity patterns are undisturbed. In addition, the SP is able to assign personalized pricing policies that strike the best balance between enhancing the certainty about the future demand for optimal proactive caching, and maximizing the revenue collected from end users. Comparing the proposed system's performance with the baseline scenario of the existing practice of no-proactive service, we show that the SP attains profit gain that grows with number of users, at least, as the first derivative of the cost function. Moreover, end-users that receive proactive caching services make strictly positive savings. Thus, we essentially demonstrate the win-win situation to be reaped through the exploitation of the consistent users' activity.**

***Index Terms*—Resource allocation, wireless networks, convex optimization, pricing.**

I. Introduction

The significant transformation of the wireless data traffic from text and voice communication to high data rate multimedia applications (e.g., HD video streaming) has raised major concerns on the ability of the limited wireless spectrum to support reliable data delivery at reasonable costs [1] for service providers (SPs), including carriers (e.g., AT&T, T-Mobile), and content sources (e.g., YouTube, CNN, Netflix). On the other hand, several bands of the wireless spectrum face a consistent underutilization issue during the off-peak time because of the lack of user access.

Thus, there has been an urgent need for more advanced and sophisticated techniques to boost the wireless resource management. The notion of Dynamic spectrum access (DSA) has been introduced to allow out-of-band (cognitive or secondary) users to access the underutilized bands of a primary network at lower prices [2]. However, it is expected that DSA will offer only a partial solution to the problem as most users already idle during the off-peak time.

This work was primarily supported by the QNRF under Grants NPRP 7-923-2-344 and NPRP 09-1168-2-455. The work of A. Eryilmaz was also supported by the NSF grants: CAREER-CNS-0953515, CNSWiFiUS-1456806, and CCSS-EARS-1444026; and the DTRA grant: HDTRA1-15-1-0003.

While DSA implicitly relies on pricing incentives to enhance the spectrum utilization, it does not offer an effective solution to the excessive demands of the peak time of the day. More research directions have been opened to tackle such a problem. In [3] time-of-the-day pricing has been developed based on the instantaneous load level of the system with prices potentially increase when the load rises and decrease otherwise. The model, however, falls short of capturing the inherent variability of the system loads due to the users activity patterns, not only pricing. In a more recent work, [4], consistency of user activity patterns have been taken into account through a joint pricing and scheduling framework that incentivizes users to delay their demand to the off-peak time through optimized discounts. The potentially reaped gains of this approach rely on the tradeoff between the economic responsiveness of the users, and the preferred time of access. Clearly, users will have to change their regular activity patterns in order to retrieve low-priced service, for example a user who wants to access the network while in public transit, may have to defer demand to later time when network prices are less, and thus gets pricing discount at the expense of inconvenient access pattern. Hence, the SP gains are not surprising.

In a parallel direction, WiFi infrastructure has been proposed to off-load the cellular traffic for wireless users with WiFi connection capabilities [5], [6]. The idea is to re-route the users demand through WiFi access points covering the users connections. While the emerging results reveal a potential of promising gains, the approach does not effectively tackle the excessive costs and low quality-of-service (QoS) in places with no WiFi coverage (e.g., public transit, rallies, stadiums, etc.). Moreover, it is questionable that large amount of traffic will be efficiently offloaded through WiFi since users by default connect to WiFi as a cheaper and more reliable source of data communication, and therefore the fraction of users who direct their demand to the cellular provider while WiFi covered is expected to be minor.

Content sources, such YouTube and Netflix, employ content distribution networks (CDNs), e.g., Akamai and Limelight, to enhance the reliability of data delivery at the peak time and provide timely service of content requests (see e.g., [7]). CDNs in turn seek to maximally utilize their storage resources through efficient content placement strategies which incorporate dynamic popularity, and cache location and size [7]-[10]. Research works in [11], [12] consider proactive scheduling for CDNs with predictable content popularity, yet demand

characteristics have been assumed identical over time, hence do not capture the the peak-to-average-ratio problem addressed in this work.

Recently, we have proposed the notion of *proactive resource allocation* [13], [14] as a remedy to the peak-to-average-ratio problem. The technique aims to exploit the *predictable* human demand together with the large storage offered by the smart wireless devices in smoothing out the wireless network traffic over time by *proactively serving peak-hour requests during the off-peak time*. Thus, users need not change their regular activity patterns, yet they will be promised better QoS. By moving caching to end-users, SPs can attain higher performance gains compared to the use of CDNs only since optimized individual proactive services can further utilize available resources. In addition, it enables personalized incentives for enhanced certainty about future demand.

We have demonstrated the potential gains of proactive resource allocation, under large-timescale optimization, for unicast and multicast networks as well as networks with heterogeneous QoS requirements under perfectly predictable demand. Further, authors in [15] characterize potential delay reduction gains of proactive scheduling under perfect predictability. In [16], [17] we have considered the impact of future demand uncertainty and have introduced the notion of demand shaping through pricing incentives. The idea is to price data content based on the user preferences in the sense that more popular content receives less prices, hence users become more attracted to it, the certainty about future demand increases, and the SP attains efficient proactive download performance. In [17], we have shown in a single-user that not only does the proposed joint proactive service and pricing schemes improve the SP's profit, but also the user pays less without changing his activity pattern over time.

In this paper, we generalize our results to the multi-user system served by a content source, and establish asymptotic bounds on the extra profit the SP can make through proactive scheduling. We develop a distributed algorithm where users can determine proactive downloads with the promise of less payment. In particular, the contributions of this paper along with its structure are as follows.

• In Section II, we introduce the system setup, model the demand profile and economic responsiveness of end users, and formulate the content based pricing problem together with proactive data download.

• In Section III, we study the single user case of the problem. We derive necessary and sufficient condition on the activity of the end user and his preferences so that the smart pricing with proactive downloads can always yield positive profit and savings gains for the SP and end-user relative to the flat pricing policy without proactive service. We develop an algorithm that attains these gains and prove its convergence.

• In Section IV, we address the multi-user case of the problem. We tackle the non-convexity of the price allocation problem through successive approximations of convex problems, and develop centralized and distributed algorithms to implement the joint allocation of pricing and proactive download policies. We show that the SP can reap profit gains that grow unboundedly with the number of users through proactive services. Moreover, all users who receive such service make positive savings gain. Thus we highlight the potential for the win-win situation offered by leveraging the predictability of human behavior.

• In Section V, we present numerical simulations to validate the theoretical results and demonstrate the system gains. We conclude the work in Section VI.

II. System Model and Problem Formulation

The system model has some similarities to those in [16], [14]. We consider a system comprising a SP and a set of N users. We consider the SP to be a content source such as YouTube, Netflix, CNN, Facebook, etc., In a time slotted fashion, users generate random requests targeting content from one of possible M data items available at the SP. The SP in turn responds to such users' requests in a timely basis. The content of a data item could be a movie (as in YouTube and Netflix), a soundtrack (as in Pandora), a news update (as in CNN and Fox News), etc.

We assume that such contents are updated consistently every time slot, where the slot duration represents the time over which a user consumes the content of requested data item. This can span a few minutes to possibly hours. Thus, in the slot duration, the SP completes delivery of requested data content, which has been partially served over the previous slot.

Users' demand profiles: At time slot t, user n may decide to consume a total of $S_{n,t}$ amount of data, where $\check{S} \leq S_{n,t} \leq \hat{S}$, $\forall n, t$, and $0 < \check{S} \leq \hat{S} < \infty$. This consumption is from one of the M data items, per slot. The user decides to consume content from item m at time t with probability $P_{n,t}(m)$, $m = 1, \cdots, M$. We denote the demand profile for user n at time t by the vector $\mathbf{p}_{n,t} = \{P_{n,t}(m)\}_{m=1}^{M}$.

Each user exhibits a different willingness-to-pay value for every data item, which potentially captures the value of such item to the user. The willingness-to-pay value for item m as recognized by user n is denoted by $v_n(m)$. We assume the SP can measure these values through different methods (see e.g., [25] and the references therein), and through the significant advances in machine learning and collaborative filtering for personalized recommendation techniques [26], [27]. As content of each data item may be updated every time slot, we assume willingness-to-pay value to vary over a larger time horizon, e.g., weeks or months. The SP assigns a price $y_{n,t}(m) \leq v_n(m)$ for data item m when it is requested by user n at time t. The impact of such price on the probability of requesting item m is captured through the mapping $P_{n,t}(m) = \phi_{m,n,t}(y_{n,t}(m))$, which is non-negative and non-decreasing in $y_{n,t}(m)$[1]. We model the statistics of the predictable user demands as follows:

- The demand of user n to item m at slot t is captured by a random variable $\mathbb{I}_{n,t}(m)$ where

$$\mathbb{I}_{n,t}(m) = \begin{cases} 1, & \text{with probability } P_{n,t}(m), \\ 0, & \text{with probability } 1 - P_{n,t}(m). \end{cases}$$

[1]While we consider a general pricing model that allows content and user dependent pricing, the notion of pricing here is not restricted to *currency*. Instead, it can be a proxy to virtual tokens or points assigned to users based on their subscriptions.

- Any two different users n, k have $\mathbb{I}_{n,t}(m)$ independent of $\mathbb{I}_{k,t}(j)$ for all m, j.
- Each user n can consume content from at most one data item per time slot[2]. Hence $\sum_{m=1}^{M} \mathbb{I}_{n,t}(m) \leq 1$.
- Any user n is idle at slot t with probability $q_{n,t} := 1 - \sum_{m=1}^{M} P_{n,t}(m)$.

Further, motivated by the consistent user activity patterns shown in [18], and the reported results on human behavioral patterns in [19], we assume that the demand profile of each user follows a *cyclostationary* pattern that repeats itself consistently in a period of T time slots. The T-slot period can be interpreted as a single day through which the activity of each user varies each hour, but occurs with the same statistics consistently each day. That is, $S_{n,t} = S_{n,t+k}$ and $q_{n,t} = q_{n,t+T}$ for any non-negative integer k, and $t = 0, 1, \cdots$.

Proactive download decisions: To balance its aggregate load over time, the SP assigns one-slot-ahead proactive download values from the potential demand of the next slot to be served in the current. In particular, the SP sets $x_{n,t+1}(m)$ as the proactive download from the content of item m to be served to user n with the demand of time t. The assumption of only one-slot ahead proactive service is to account for content freshness guarantees so that cached content is relevant to the user at the time of consumption.

Incurred cost: The SP incurs costs proportional to the aggregate load at each slot. In particular, if L is the total SP load at a specific slot, then the cost incurred in such slot is $C(L)$, for some smooth, strictly convex, and monotonically increasing cost function $C : \mathbb{R}_+ \to \mathbb{R}_+$. Denoting by $\mathbf{x}$ and $\mathbf{y}$, respectively, the collections of proactive downloads and service prices for all users, time slots, and data items, the SP's time average expected cost, thus, can be written as

$$\eta(\mathbf{x}, \mathbf{y}) := \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[C\left(\sum_{m=1}^{M} \sum_{n=1}^{N} (S_{n,t} - x_{n,t}(m)) \mathbb{I}_{n,t}(m) + x_{n,t+1}(m) \right) \right]. \quad (1)$$

Please note the contribution of the proactive download $x_{n,t+1}(m)$ to the total load at time t, which translates to a cost that the SP incurs.

Revenue: The SP also reaps revenue due to the users payments for the requested content. Such revenue essentially depends on the users' demand profiles which are shaped through the pricing policy used. The time average revenue (or user payments) received by the SPs is given by

$$\mu(\mathbf{y}) := \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \sum_{n=1}^{N} y_{n,t}(m) \phi_{n,m,t}(y_{n,t}(m)). \quad (2)$$

As such, the SP's profit is given by $R(\mathbf{x}, \mathbf{y}) := \mu(\mathbf{y}) - \eta(\mathbf{x}, \mathbf{y})$. The ultimate objective is to find the optimal allocation of $\mathbf{x}$ and $\mathbf{y}$ that maximize such a profit. The time average profit maximization to be solved by the SP is formulated as

$$R^* := \max_{\mathbf{x},\mathbf{y}} \quad R(\mathbf{x}, \mathbf{y}) \quad (3)$$

$$\text{subject to,} \quad x_{n,T}(m) = x_{n,0}(m), \forall m, n, \quad (4)$$

$$0 \leq x_{n,t}(m) \leq S_{n,t}, \quad \forall m, n, t, \quad (5)$$

$$0 \leq y_{n,t}(m) \leq v_n(m), \quad \forall m, n, \quad (6)$$

$$\sum_{m=1}^{M} \phi_{m,n,t}(y_{n,t}(m)) = 1 - q_{n,t}, \quad \forall n, t. \quad (7)$$

Here, the activity of each user over time is kept unchanged through the constraints $\sum_{m=1}^{M} \phi_{m,n,t}(y_{n,t}(m)) = 1 - q_{n,t}$ which ensures same access probability.

Here we note that the above formulation can also apply to the T-slot finite horizon optimization of the problem by replacing (4) with $x_{n,0}(m) = 0, \forall m, n$. Such finite horizon scenario can be adopted when the SP is unaware of the demand characteristics beyond time T.

In the formulation above, there are convexity issues with the following. (i) The objective function is not concave in $(\mathbf{x}, \mathbf{y})$ due to the product form of the cost as a function of proactive downloads and probability of demand as a function of price. (ii) The functions $\phi_{m,n,t}(y_{n,t}(m))$ may be non-affine in $y_{n,t}(m)$, hence yielding non-convex constraint set [20].

Yet, the studies carried out in [4] have revealed that users exhibit *linear* responsiveness to pricing incentives, which we will consider in our main investigations throughout the rest of the paper. In particular, we will adopt

$$\phi_{m,n,t}(y_{n,t}(m)) := \frac{1 - q_{n,t}}{D_{n,t}} (v_n(m) - y_{n,t}(m)), \quad \forall m, n, t,$$

where $D_{n,t}$ is a normalizing constant. Further, we will provide remarks along with the main results to cover the case when $\phi_{m,n,t}$ is not an affine function. With the above form for $\phi_{m,n,t}$, constraints (7), reduce to an average price constraint

$$\frac{1}{M} \sum_{m=1}^{M} y_{n,t}(m) = \bar{y}_{n,t}, \quad \forall n, t, \quad (8)$$

where $\bar{y}_{n,t} := \sum_m v_n(m) - D_{n,t}$.

Our main focus in the following sections is to tackle the inherent non-convexity issue of the objective function, and show the system gains under suboptimal, yet efficient, algorithms. We note that, in (3), we assume the SP has *only* statistical knowledge about the users activity, which is justified since most SPs consistently log and track user interactions with the network over time and can construct the needed demand profiles through the aid of machine learning tools such as collaborative filtering. Hence, in the sequel we focus on the design of off-line algorithms which take place as a one-shot that determines pricing and proactive caching for every slot. In the next section, we begin with the single user scenario for its simplicity, then generalize the design to the multi-user case in the subsequent section.

III. The Single-user Scenario

In this subsection, we consider $N = 1$, thus we omit the subscript n from all the notation. While the single user scenario facilitates an ease of exposition, it also fits the setup

[2]Please note that a *data item* is not restricted to contain only one video, as for example one can consider CNN as a data item with multiple new videos update it every time slot. A user therefore can consume all these videos in one slot, while another user consumes only one video with roughly the same duration from a YouTube channel which can be treated as another data item.

whereby the contribution of the users' requests to the aggregate cost is not coupled by the convex cost function. In this case, the proactive download and price allocation problem (3) reduces to

$$R^* := \max_{\{\mathbf{x},\mathbf{y}\}_{t=0}^{T-1}} \quad \frac{1}{T}\sum_{t=0}^{T-1}\frac{1-q_t}{D_t}\left[\sum_{m=1}^{M}(v(m)-y_t(m))\left(y_t(m) - C\left(S_t - x_t(m) + \sum_{j=1}^{M} x_{t+1}(j)\right)\right)\right] - q_t C\left(\sum_{j=1}^{M} x_{t+1}(j)\right) \tag{9}$$

$$\text{subject to, constraints } (4),(5),(6),(8), \text{with } N=1.$$

It can be noted from (9) that the non-convexity of the objective function is still restraining the achievability of the globally optimal solution. However, an efficient solution to it can still be attained as described in the following subsection.

A. Single-user Iterative Pricing and Proactive Download

Our proposed suboptimal algorithm takes an iterative sequential form of price allocation followed by proactive download steps. In particular, we start the algorithm with some initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, then, at every iteration k, the algorithm goes through the following two steps:

- **Step 1: Profit maximization:**

$$R^{(k)} := \max_{\mathbf{y}\in\mathcal{Y}} \quad R(\mathbf{x}^{(k-1)}, \mathbf{y}) \tag{10}$$

where $\mathbf{x}^{(k-1)}$ is the optimal solution of **Step 2** at iteration $k-1$,

$$\mathcal{Y} := \left\{\mathbf{y} : 0 \le y_t(m) \le v(m), \quad \frac{1}{M}\sum_{m=1}^{M} y_t(m) = \bar{y}_t, \forall m, t\right\}.$$

- **Step 2: Cost minimization:**

$$\eta^{(k)} := \min_{\mathbf{x}\in\mathcal{X}} \quad \eta(\mathbf{x}, \mathbf{y}^k) \tag{11}$$

where $\mathbf{y}^{(k)}$ is the optimal solution of **Step 1** at iteration k,

$$\mathcal{X} := \{\mathbf{x} : x_T(m) = x_0(m), \quad 0 \le x_t(m) \le S_t, \quad \forall m, t\}.$$

The algorithm utilizes the observation that the profit maximization problem given a proactive download control $\mathbf{x}$ is a convex function in $\mathbf{y}$. Moreover, given a price allocation strategy $\mathbf{y}$, the profit function is convex in $\mathbf{x}$, and its maximization is equivalent to minimizing the expected cost.

The following theorem establishes the convergence of the proposed suboptimal algorithm.

Theorem 1: The sequence of SP profits $\{R^{(k)}\}_k$ generated through the iterative solutions of (10), (11) is increasing and convergent.

Proof. We have for any iteration k,

$$\begin{aligned} R^{(k)} &= \max_{\mathbf{y}\in\mathcal{Y}} R(\mathbf{x}^{(k-1)}, \mathbf{y}) \ge R(\mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)}) \\ &\ge R(\mathbf{x}^{(k-2)}, \mathbf{y}^{(k-1)}) = R^{(k-1)} \end{aligned}$$

where the last inequality follows since $x^{(k-1)} := \arg\min_{\mathbf{x}} \eta(\mathbf{x}, \mathbf{y}^{(k-1)}) = \arg\min_{\mathbf{x}\in\mathcal{Y}} R(\mathbf{x}, \mathbf{y}^{(k-1)})$.

Thus, $R^{(k)} \ge R^{(k-1)}$, $\forall k \ge 0$. Now, since $R^{(k)}$ is bounded above by $R^{(k)} \le \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} v(m) < \infty$, then it converges to a finite limit $R^* := R(\mathbf{x}^*, \mathbf{y}^*)$ with $\mathbf{x}^*, \mathbf{y}^*$ are the limit points of $\{\mathbf{x}^{(k)}\}_k$,$\{\mathbf{y}^{(k)}\}_k$, respectively. ■

Thus, starting from any initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, the algorithm converges to a point $(\mathbf{x}^*, \mathbf{y}^*)$ for which $R(\mathbf{x}^*, \mathbf{y}^*) \ge R(\mathbf{x}^{(0)}, \mathbf{y}^{(y)})$. In the following subsection, we investigate the performance of the algorithm in terms of SP profit and end user payments.

B. Performance Analysis

For performance evaluation, we compare the results of the developed algorithm in the previous subsection with a baseline profit and user expected payment in which no proactive downloads are scheduled. In other words, the proactive download control of the baseline scenario is $\hat{\mathbf{x}} := \mathbf{0}$, and the associated pricing policy is $\hat{\mathbf{y}} := \arg\max_{\mathbf{y}\in\mathcal{Y}} R(\mathbf{y}, \mathbf{0})$. The expected profit, cost, and user payments under the baseline scenario are denoted respectively by $\hat{R}$, $\hat{\eta}$, and $\hat{\mu}$.

In our analysis we focus on the performance of the proposed algorithm when the initial point is of the baseline scenario, i.e., $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) = (\mathbf{0}, \hat{\mathbf{y}})$. The idea behind such a choice is to highlight the potential gains to be reaped by such system when proactive data download is employed. In fact, under the baseline scenario, the SP can assign time and data item dependent prices to maximize its profit. However, with the additional degree of freedom of proactive service, such prices can be optimized to realize more gains.

Letting $(\mathbf{x}^*, \mathbf{y}^*)$ denote the limit point of the proposed algorithm when the initial point is $(\mathbf{0}, \hat{\mathbf{y}})$, we define $R^* := R(\mathbf{x}^*, \mathbf{y}^*)$, $\eta^* = \eta(\mathbf{x}^*, \mathbf{y}^*)$, and $\mu^* = \mu(\mathbf{y}^*)$. We compare the profit and savings gains of the proposed algorithm relative to the baseline through the quantities: profit gain $\Delta R := R^* - \hat{R}$, savings gain $\Delta\mu := \hat{\mu} - \mu^*$, and cost reduction gain $\Delta\eta := \hat{\eta} - \eta^*$. Clearly, we have $\Delta\eta = \Delta R + \Delta\mu$. In other words, the cost reduction achieved through proactive downloads is divided over the profit gain and savings gain.

In the following theorem, we establish a lower bound on the profit gain.

Theorem 2: Let $m_t^* := \arg\max_{m\in\{1,\cdots,M\}} \phi_{m,t}(\hat{y}_t(m))$, and $\mathcal{T} := \left\{t : \phi_{m_t^*,t}(\hat{y}_t(m_t^*))C'(S_t) > (1-q_{t-1})C'(S_{t-1})\right\}$, $\forall t$. Then,

$$\Delta R \ge \frac{1}{T}\sum_{t\in\mathcal{T}} \chi_t \left(\phi_{m_t^*,t}(\hat{y}_t(m_t^*))C'(S_t - \chi_t) - \mathbb{E}\left[C'\left(\sum_{j=1}^{M} S_{t-1}\hat{\mathbb{I}}_{t-1}(j) + \chi_t\right)\right]\right), \tag{12}$$

where C' is the first derivative of C,

$$\chi_t := \arg\max_{0\le x\le S_t} x\left(\phi_{m_t^*,t}(\hat{y}_t(m_t^*))C'(S_t - x) - \mathbb{E}\left[C'\left(\sum_{j=1}^{M} S_{t-1}\hat{\mathbb{I}}_{t-1}(j) + x\right)\right]\right),$$

and $\hat{\mathbb{I}}_t(m)$ is the random variable capturing the event that the user requests item m at time t under pricing policy $\hat{\mathbf{y}}$.

Proof. By the monotonicity of $\{R^{(k)}\}_k$, we have $\Delta R \ge R(x^{(1)}, \hat{\mathbf{y}}) - R(\mathbf{0}, \hat{\mathbf{y}})$, i.e. $\Delta R \ge \eta(\mathbf{0}, \hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)}, \hat{\mathbf{y}})$. Further, suppose we use a suboptimal proactive download strategy

rather than $\mathbf{x}^{(1)}$. In such a strategy, we set $x_t(m) = \chi_t$ if $m = m_t^*$, and $t \in \mathcal{T}$, and $x_t(m) = 0$, otherwise. Hence,

$$\begin{aligned}\Delta R &\geq \eta(\mathbf{0}, \hat{\mathbf{y}}) - \eta(\mathbf{x}, \hat{\mathbf{y}}) \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[C\left(\sum_{m=1}^{M} S_t \hat{\mathbb{I}}_t(m)\right)\right] - \\ &\mathbb{E}\left[C\left(\sum_{m=1}^{M}(S_t - x_t(m))\hat{\mathbb{I}}_t(m) + x_{t+1}(m)\right)\right] \\ &\overset{(a)}{\geq} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[C'\left(\sum_{m=1}^{M}(S_t - x_t(m))\hat{\mathbb{I}}_t(m) + x_{t+1}(m)\right)\times\right. \\ &\left.\left(\sum_{m=1}^{M} x_t(m)\hat{\mathbb{I}}_t(m) - x_{t+1}(m)\right)\right] \\ &\overset{(b)}{=} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} x_t(m)\mathbb{E}\left[\hat{\mathbb{I}}_t(m)C'\Big(\sum_{j=1}^{M}(S_t - x_t(j))\hat{\mathbb{I}}_t(j)\right. \\ &\left. + x_{t+1}(j)\Big) - C'\left(\sum_{j=1}^{M}(S_{t-1} - x_{t-1}(j))\hat{\mathbb{I}}_{t-1}(j) + x_t(j)\right)\right] \\ &\overset{(c)}{\geq} \frac{1}{T}\sum_{t\in\mathcal{T}}\chi_t\left(\phi_{m_t^*,t}(\hat{y}_t(m_t^*))C'(S_t - \chi_t) - \right. \\ &\left.\mathbb{E}\left[C'\left(\sum_{j=1}^{M} S_{t-1}\mathbb{I}_{t-1}(j) + \chi_t\right)\right]\right),\end{aligned}$$

Inequality (a) follows by the mean value theorem for random variables [21] while noting that C' is monotonically increasing and non-negative. Equality (b) follows by rearranging the terms. Inequality (c) follows since we omit the non-negative values of $x_{t+1}(j)$ and $x_{t-1}(j)$, $\forall j$ which can only reduce the right hand side of Inequality (b). ■

The established bound in Theorem 2 essentially relies on the existence of a sufficient difference between the marginal costs of two consecutive slots, clearly, if $\mathcal{T}$ is non-empty, then the lower bound is strictly positive. For instance, the user may be busy at work until the end of the business day, where he decides to follow up the news and request data content. The SP can therefore partially supply such data before the end of the business day so as to regulate the traffic over time. We also note that, if $\mathcal{T} = \{\}$, then $\Delta R = 0$, which potentially captures the event of a user with balanced demand over time, for which the SP does not gain by shifting the load in time.

Now, we study the effect of proactive downloads on the savings gain, $\Delta\mu$. We first establish the statement that proactive downloads can not yield a larger expected payment than the baseline scenario. Subsequently, we characterize the necessary and sufficient conditions for strictly positive savings gain.

Lemma 1: The expected payment by the user under the proposed iterative algorithm satisfies $\mu(\mathbf{y}^*) \leq \mu(\hat{\mathbf{y}})$.

Proof. Since $R(\mathbf{0}, \hat{\mathbf{y}}) = \max_{\mathbf{y}\in\mathcal{Y}} \mu(\mathbf{0}, \mathbf{y}) - \eta(\mathbf{0}, \mathbf{y})$ with $\eta(\mathbf{0}, \mathbf{y}) = \sum_{t,m} C(S_t)\phi_{m,t}(y_t(m))$, the constraint that $\sum_m \phi_{m,t}(y_t(m)) = 1 - q_t$ implies $\eta(\mathbf{0}, \mathbf{y}) = \sum_t C(S_t)(1 - q_t)$ is independent of $\mathbf{y}$. In this case, profit maximization will reduce to maximizing the expected payment. Therefore, $\mu(\mathbf{0}, \hat{\mathbf{y}})$ is the maximum expected payment by the user. ■

The user will be charged the maximum possible payment when the SP can not modify its load dynamics over time. With the proactive downloads, however, the smoothed-out load characteristics will never yield extra payments on the user. The following theorem characterizes an even stronger result.

Theorem 3: Suppose that $C(0) = 0$, and $\max_m v(m) > \min_m v(m)$, then the savings gain satisfies $\Delta\mu > 0$ if and only if

$$\phi_{m_0,t_0}(\hat{y}_{t_0}(m_0))C'(S_{t_0}) > (1 - q_{t_0-1})C'(S_{t_0-1}), \qquad (13)$$

for some time slot t_0, and data item m_0.

Proof. Since $\max_m v(m) > \min_m v(m)$, then $\hat{y}_{t_0}(m_1) \neq \hat{y}_{t_0}(m_2)$ for some two items m_1, m_2. Hence, if ${x^*}_{t_0}(m_0) > 0$, we get $\max_m {x^*}_{t_0}(m) > \min_m {x^*}_{t_0}(m)$, which eventually yields $\mu(\mathbf{y}^*) < \mu(\mathbf{0}, \hat{\mathbf{y}})$. Now, we show that (13) is necessary and sufficient to have ${x^*}_{t_0}(m_0) > 0$.

($\Rightarrow$) Suppose that (13) holds. We set $x_t(m) = 0$ for all $(m, t) \neq (m_0, t_0)$. Consider the difference between the expected costs (we adopt $C(0) = 0$),

$$\begin{aligned}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\left(C(S_t) - C\left(S_t - x_t(m) + \sum_j x_{t+1}(j)\right)\right)\phi_{m,t}(\hat{y}_t(m)) \\ -C\left(\sum_j x_{t+1}(j)\right)q_t = (C(S_t) - C(S_t - x_{t_0}(m_0)))\times \\ \phi_{m_0,t_0}(\hat{y}_{t_0}(m)) + (C(S_t) - C(S_t + x_{t_0}(m_0)))(1 - q_{t_0-1}) \\ +(C(0) - C(x_{t_0}(m_0)))q_{t_0-1} \\ \geq x_{t_0}(m_0)G(x_{t_0}(m_0)),\end{aligned}$$

where

$$\begin{aligned}G(x_{t_0}(m_0)) := C'(S_t - x_{t_0}(m_0))\phi_{m_0,t_0}(\hat{y}_{t_0}(m_0)) - \\ C'(x_{t_0}(m_0))q_{t_0-1} - C'(S_{t-1} + x_{t_0}(m_0))(1 - q_{t_0-1}).\end{aligned}$$

The last inequality follows by the mean value theorem.

Since G is decreasing in $x_{t_0}(m_0)$, and $G(0) = C'(S_{t_0})\phi_{m_0,t_0}(\hat{y}_{t_0}(m_0)) - C'(S_{t_0-1})(1 - q_{t_0-1}) > 0$ by hypothesis. By the continuity of C', there exists $x > 0$ for which $xG(x) > 0$. That is, $x_{t_0}(m_0) > 0$ will yield a strictly reduced cost as compared to the no proactive download scenario.

($\Leftarrow$) Suppose that $\mathbf{x}^* > 0$. Suppose towards contradiction that (13) does not hold. Since $\mathbf{x}^* > 0$, the cost reduction due to proactive service can only be upper bounded by a positive value. That is,

$$\begin{aligned}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\left(C(S_t) - C\left(S_t - \hat{x}_t(m) + \sum_j \hat{x}_{t+1}(j)\right)\right)\phi_{m,t}(\hat{y}_t(m)) \\ -C\left(\sum_j x^*_{t+1}(j)\right)q_t \leq \sum_{t=0}^{T-1}\sum_{m=1}^{M} C'(S_t)\times \\ \left(x^*_t(m) - \sum_j x^*_{t+1}(j)\right)\phi_{m,t}(\hat{y}_t(m)) = \\ \sum_{t=0}^{T-1}\sum_{m=1}^{M} x^*_t(m)(\phi_{m,t}(\hat{y}_t(m))C'(S_t) - (1 - q_{t-1})C'(S_{t-1})),\end{aligned}$$

hence if (13) does not hold for any m or t, the right hand side of the last expression will be non-positive, which contradicts the hypothesis that $\mathbf{x}^* > 0$, with positive cost reduction. Therefore (13) must hold. ■

Remark 1: In Theorem 3, suppose that $\min_m v(m) \geq |2\bar{y}_{t_0} - \bar{v}|$, where $\bar{v} = \frac{1}{M}\sum_m v(m)$. Then Condition (13)

reduces to

$$\frac{1}{M}+\frac{1}{2}\cdot\frac{\max_m v(m)-\bar{v}}{D_{t_0}}>\frac{(1-q_{t_0-1})C(S_{t_0-1})}{(1-q_{t_0})C(S_{t_0})}. \quad (14)$$

The requirement on $\min_m v(m)$ facilitates a full characterization of $\hat{\mathbf{y}}$ as a function of $(v(m))_m$ and D_t.

The left hand side of (14) is the probability of requesting the service with the maximum willingness-to-pay value, given that the user is going to request a service at time t_0. Here, $\hat{y}_{t_0}(m)=\frac{v_m+\bar{v}}{2}-\frac{D_{t_0}}{M}$. Thus, it is required to have this probability greater than the ratio between its activity at the previous slot t_0-1 over the activity at t_0.

Remark 2: In the proofs of Theorems 1, 2, 3, we used a general form for $\phi_{m,t}$ which is not necessarily linear in $y_t(m)$. This therefore extends the results of those theorems to any algorithm that yields a locally optimal solution to (9) under general economic responsiveness.

C. Distributed Implementation

As we have seen that the end-user has a potential to achieve positive savings gain through proactive downloads, the SP can involve the user to carryout the proactive download decisions of the proposed iterative algorithm in response to the user the assigned prices. In particular, the SP can run **Step 1** of the algorithm and assign a certain pricing policy that maximizes the profit given the adopted proactive download strategy by the end-user.

The end-user reacts to such pricing policy by executing **Step 2** of the algorithm so as to balance its load over time, reduce the cost at the SP's end, and essentially attain less payments. The two parties will iterate their respective steps until convergence. From a game-theoretic perspective, the point $(\mathbf{x}^*,\mathbf{y}^*)$ to which the algorithm converges is a Nash equilibrium point of the dynamic coordinated game incorporating the SP and end-user as its players, with the common objective of maximizing the profit gain [22].

IV. The Multi-user Scenario

The insights derived from the single-user scenario will inspire the development of an iterative algorithm that essentially yields a similar win-win situation to SPs and end-users who employ proactive content download. In the iterative single-user algorithm, we have observed that the separation of the price allocation and proactive download decisions leads to two convex problems which can be solved iteratively until convergence. In the multi-user case, the price allocation step suffers from a product of prices for many users that need to be jointly optimized, thus rendering the price allocation problem non-convex. In this section, we tackle such an additional difficulty through tools from non-convex optimization [24], and establish performance bounds as in the previous section.

A. Multi-user Iterative Pricing and Proactive Download

We consider the sequence of solutions to approximate convex problems of (3). In particular, in each iteration k, we replace the objective function $R(\mathbf{x},\mathbf{y})$, which is non-concave, with an approximate concave function $\tilde{R}^{(k)}(\mathbf{x},\mathbf{y})$, and we use the solution to the resulting convex problem to generate a new approximate function $\tilde{R}^{(k+1)}$, and so on. The series of solutions to these approximate convex problems converges to a point that satisfies the KKT conditions of (3).

Lemma 2: Let $\tilde{R}^{(k)}$ be a concave function in $(\mathbf{x},\mathbf{y})$ that replaces the objective function R of (3) at iteration k. Denote by $(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})$ the optimal solution to the resulting convex optimization problem at the $(k-1)^{st}$ iteration, $k=1,2,\cdots$. If

1) $\tilde{R}^{(k)}(\mathbf{x},\mathbf{y})\leq R(\mathbf{x},\mathbf{y})$ for all feasible $(\mathbf{x},\mathbf{y})$,
2) $\nabla\tilde{R}^{(k)}(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})=\nabla R(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})$,
3) $\tilde{R}^{(k)}(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})=R(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})$,

then $R(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})<R(\mathbf{x}^{(k)},\mathbf{y}^{(k)}),\forall k$, and the sequence $\{(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\}_k$ converges to a point $(\mathbf{x}^*,\mathbf{y}^*)$ which is a locally optimal solution to (3).

The above lemma is a special case of Theorem 1 in [24] which aims to provide local optimal solutions to non-convex problems.

Corollary 1: Starting from the baseline initial point $(\mathbf{x}^(0),\mathbf{y}^(0))=(\mathbf{0},\hat{\mathbf{y}})$, a sequence of approximate functions $\{\tilde{R}^{(k)}\}$ generated as in Lemma 2 and resulting in a KKT-satisfying point $(\mathbf{x}^*,\mathbf{y}^*)$ leads to $R(\mathbf{0},\tilde{\mathbf{y}})<f_0(\hat{\mathbf{x}},\hat{\mathbf{y}})$.
Thus, the resulting local optimal solution will essentially lead to a better profit performance than that of the baseline scenario when no-proactive downloads are carried out.

In the following theorem, we provide a particular approximation to R of (3) at each new iteration k that satisfies the requirements of Lemma 2.

Theorem 4: For R being the objective function of (3), the approximate function

$$\tilde{R}^{(k)}(\mathbf{x},\mathbf{y})=\mu(\mathbf{y})-\eta(\mathbf{x},\mathbf{y}^{(k-1)})-\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m,n}h_{n,t}^{(k)}(m)y_{n,t}(m), \quad (15)$$

where

$$h_{n,t}^{(k)}(m)=\left.\frac{\partial\eta(\mathbf{x}^{(k-1)},\mathbf{y})}{\partial y_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})}, \quad \forall m,n,t, \quad (16)$$

at iteration $k\geq 1$ is concave in $(\mathbf{x},\mathbf{y})$. Further, the sequence of solutions to the problem resulting from replacing R with $\{\tilde{R}^{(k)}\}_k$ converges to a locally optimal solution of (3).

Proof. Please refer to Appendix A. ■

With the construction of the approximate functions $\tilde{R}^{(k)}$ in the above theorem, we now present the iterative price and proactive download algorithm.

- Start with some initial point $(\mathbf{x}^{(0)},\mathbf{y}^{(0)})$.
- At each iteration $k\geq 1$, compute $\tilde{R}^{(k)}$ as in (15), and solve the convex problem

$$\max_{\{\mathbf{x},\mathbf{y}\}_{m,n,t=0}^{T-1}} \tilde{R}^{(k)}(\mathbf{x},\mathbf{y}) \quad (17)$$

$$\text{subject to , constraints } (4),(5),(6),(8). \quad (18)$$

In the following subsection, we focus on the performance analysis of the proposed multi-user iterative algorithm.

B. Performance Analysis

To quantify the performance of the multiuser iterative algorithm, we utilize the same metrics of the single-user counterpart. In particular, we take the initial point of the algorithm as the baseline no-proactive download scenario. That is, $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) = (\mathbf{0}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} R(\mathbf{0}, \mathbf{y})$, and here we take the set $\mathcal{Y}$ as

$$\mathcal{Y} := \Bigg\{ \mathbf{y} : 0 \leq y_{n,t}(m) \leq v_n(m), \sum_{m=1}^{M} \phi_{m,n,t}(y_{n,t}(m)) = 1 - q_{n,t}, \forall m, n, t \Bigg\}.$$

We use the notation $(\mathbf{x}^*, \mathbf{y}^*)$ to capture the point to which the sequence of solutions $\{(\mathbf{x}^{(k)}), \mathbf{y}^{(k)}\}_k$ generated according to the multi-user iterative algorithm converges. Further we use ΔR, $\Delta\mu$, and $\Delta\eta$ to quantify the profit, savings, and cost reduction gains, respectively, where $\Delta R = R(\mathbf{x}^*, \mathbf{y}^*) - R(\mathbf{0}, \hat{\mathbf{y}})$, $\Delta\mu = \mu(\hat{\mathbf{y}}) - \mu(\mathbf{y}^*)$, and $\Delta\eta = \eta(\mathbf{0}, \hat{\mathbf{y}}) - \eta(\mathbf{x}^*, \mathbf{y}^*)$.

To characterize a lower bound on the profit gain, we introduce the following notion of active users:

Definition 1 (Active users): For each data item m and time slot t, we define a set $\mathcal{B}_t(m)$ of active users as

$$\mathcal{B}_t(m) := \Big\{ n : \mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m) C'(\hat{L}_t) - C'(\hat{L}_{t-1})\right] > 0 \Big\}, \quad t = 0, \cdots, T-1, \quad m = 1, \cdots, M,$$

where $\hat{L}_t := \sum_{m,n} S_{n,t} \hat{\mathbb{I}}_{n,t}(m)$. We also use $B_t(m) := |\mathcal{B}_t(m)|$ is the cardinality of set $\mathcal{B}_t(m)$.

The active users of slot t w.r.t. item m are those users who, in the expected sense, create higher marginal cost in slot t by requesting item m than the marginal cost of the previous slot. Thus, they have a high potential to improve the cost reduction through a proactive service of their demand. We will show in the sequel that existence of such active users is necessary and sufficient for positive system gains. In the definition above, the expectation $\mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m) C'(\hat{L}_t) - C'(\hat{L}_{t-1})\right]$ captures the *marginal* contribution of user n to the cost of time slot t, when requests item m, over the cost of the previous time slot $t-1$. This, for instance, may attribute to the disparate activity levels of the users between times $t-1$ and t.

Now, we establish the following lower bound on the profit gain.

Theorem 5 (Lower bound on profit gain): For the sets of active users $\mathcal{B}_t(m)$ defined above,

$$\Delta R \geq \frac{1}{T} \sum_{t=0}^{T-1} \tilde{x}_t \sum_{m=1}^{M} \sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\Bigg[\hat{\mathbb{I}}_{n,t}(m) C'\Bigg(\hat{L}_t - \sum_{j=1}^{M} \sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t \hat{\mathbb{I}}_{k,t}(j) \Bigg) - C'\Bigg(\hat{L}_{t-1} + \sum_{j=1}^{M} \sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t \Bigg) \Bigg] > 0, \tag{19}$$

where

$$\tilde{x}_t = \arg\max_{0\leq\tilde{x}\leq\check{S}} \tilde{x} \sum_{m=1}^{M} \sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\Bigg[\hat{\mathbb{I}}_{n,t}(m) C'\Bigg(\hat{L}_t - \sum_{j=1}^{M} \sum_{k\in\mathcal{B}_t(j)} \tilde{x} \hat{\mathbb{I}}_{k,t}(j) \Bigg) - C'\Bigg(\hat{L}_{t-1} + \sum_{j=1}^{M} \sum_{k\in\mathcal{B}_t(j)} \tilde{x} \Bigg) \Bigg]$$

Proof. Please refer to Appendix B. ■

The above lower bound is achieved by allocating equal proactive download values to all active users of a certain time slot, and unchanging the pricing policy. While such a scheme is strictly suboptimal, it reveals insights on the users, data items, and time slots that can essentially benefit the system, through proactive content download, beyond the baseline.

We further consider the asymptotic behavior of the lower bound (19) when the number of users N grows to infinity. The idea is to study the impact of the number of users (and active users) on the potential profit gain when the user base expands significantly as in large rallies, stadiums, supermarkets, public transit, etc.

To that end, we start our investigations with the assumption that $q_{n,t} < 1 - \epsilon$ for all n, t and some $\epsilon > 0$. That is, each user can request a data item at any time slot with a positive probability, and this probability will not vanish as the number of users grows to infinity. Hence $\hat{L}_t \to \infty$ almost surely as $N \to \infty$, $\forall t$.

We also capture the scaling of the number of active users with N through two main assumptions. In the first, we consider the case when such a number grows to infinity with N, whereas, in the second, we consider the case when the number of active users is finite and does not scale with N.

Assumption 1 (Scaling of the number of active users): Assume that there exists some non-decreasing function $g : \mathbb{N} \to \mathbb{N}$ such that $g(N) \to \infty$ as $N \to \infty$, and for every time slot t and data item m, the limit

$$\beta_t(m) := \lim_{N\to\infty} \frac{B_t(m)}{g(N)} \tag{20}$$

exists, and $\beta_t(m) < \infty$, $\forall m = 1, \cdots, M, t = 0, \cdots, T-1$. Thus, the function $g(N)$ captures the maximum possible scaling of the active users of any item m, and time slot t with the total number of users N.

We define $\check{P}_{m,t} := \inf_{n\in\mathcal{B}_t}\{\phi_{m,n,t}(\hat{y}_{n,t}(m))\}$.

Theorem 6 (Asymptotic profit gain for infinite number of active users): Under Assumption 1, suppose that the cost function C satisfies

$$\lim_{L\to\infty} \frac{L^\delta}{C'(L)} = 0, \quad \text{for some } \delta > 0. \tag{21}$$

Suppose also that $\beta_t(m) > 0$ for some data item m and time slot t, with

$$\liminf_{N\to\infty} \{\check{P}_{m,t}\} > \limsup_{N\to\infty} \frac{\mathbb{E}[L_{t-1}]}{\mathbb{E}[L_t]}. \tag{22}$$

Then, $\Delta R(N) = \Omega(g(N) C'(\gamma N))$, for some $\gamma > 0$. In other words, the profit gain grows with N at least as $g(N) C'(\gamma N)$.

Proof. Please refer to Appendix C. ■

Note that, the assumption on $\beta_t(m) > 0$ for some m, t ensures that the number of active users grows to infinity as $g(N)$ with N. Also, Condition (22) requires the least possible contribution of an active user to the system not to vanish as the number of users grows to infinity. It can be noted from Theorem 6 that the achieved profit gain depends on the growth of the *active users* of the system with the total number of users. These are the users that essentially allow for more gains by receiving proactive service to smooth out the network load over time, and reduce the incurred costs. The first derivative of the cost function captures the rate of increase in the incurred cost as the number of users grows. Yet, it also factors-in in the lower bound demonstrating that proactive service enable superlinearly increasing profit gain with the number of users.

Remark 3 (Finite number of active users): If $0 < B_t(m) \leq B$ for all m, t, and some positive integer B, then the reaped profit gain grows with N as $\Omega(C'(\gamma N))$. That is, even under a finite number of active users, the SP is able to reap a profit grain that grows unboundedly with the number of users as $C'(\gamma N)$.

Now, we shed light on the potential savings gain $\Delta\mu$ for the users. To consider the savings of a particular user, we denote the expected payment for user n by $\mu_n(\mathbf{y})$, where

$$\mu_n(\mathbf{y}) = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} y_{n,t}(m)\phi_{m,n,t}(y_{n,t}(m))$$

and we consider the savings gain for user n as $\Delta\mu_n := \mu_n(\hat{\mathbf{y}}) - \mu_n(\mathbf{y})$.

Lemma 3: Under the proposed multi-user algorithm, the savings gain for each user n satisfies $\Delta\mu_n \geq 0$.

Proof. Starting from the initial point $(\mathbf{0}, \hat{\mathbf{y}})$, we have

$$R(\mathbf{0}, \hat{\mathbf{y}}) = \max_{\mathbf{y}\in\mathcal{Y}} \mu(\mathbf{y}) - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[C\left(\sum_{n=1}^{N} S_{n,t}\sum_{m=1}^{M}\mathbb{I}_{n,t}(m)\right)\right]$$

But $\sum_{m=1}^{M}\mathbb{I}_{n,t}(m)$ is a Bernoulli random variable with parameter $1-q_{n,t}$, i.e., independent of $\{y_{n,t}(m)\}_m$, for all n, t. Thus, $\hat{\mathbf{y}} := \arg\max_{\mathbf{y}\in\mathcal{Y}} \mu(\mathbf{y})$. Now, since $\mu(\mathbf{y})$ can be decoupled to the superposition of per-user payments $\mu(\mathbf{y}) = \sum_{n=1}^{N}\mu_n(\mathbf{y})$, it turns out that $\mu_n(\hat{\mathbf{y}}) \geq \mu_n(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$. ■

Hence, users will not pay more than their original subscription fees under no proactive downloads. However, unlike in the single user scenario, where we characterized in Theorem 3 a necessary and sufficient condition that depends only on the initial prices and willingness-to-pay values to have a positive savings gain, in the multi-user case, the dependencies of proactive download decisions of one user on the pricing policies for the others renders the derivations of similar conditions for the multi-user case analytically intractable.

Nevertheless, we can still draw some insights on the users with positive savings gain.

Theorem 7: Suppose that $\max_m v_{n_0}(m) > \min_m v_{n_0}(m)$, for some user n_0. Then $\Delta\mu_{n_0} > 0$ if and only if $x^*_{n_0,t_0}(m_0) > 0$ for some data item m_0 and time slot t_0.

Proof. The condition $\max_m v_{n_0}(m) > \min_m v_{n_0}(m)$ ensures $\{y^*_{n_0,t}(m)\}$ are not identical, hence $\{x^*_{n_0,t_0}(m)\}_m$ are also non identical if $x^*_{n_0,t_0}(m_0) > 0$.

($\Rightarrow$) Suppose that $x^*_{n_0,t_0}(m_0) > 0$. Under the multi-user iterative algorithm, since $(\mathbf{x}^*, \mathbf{y}^*)$ is the limit point of the sequence of solutions generated by the algorithm, then we can write $\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}}\mu(\mathbf{y}) - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\sum_{m=1}^{M} h_{n,t}(m)y_{n,t}(m)$

$$= \arg\max_{\mathbf{y}\in\mathcal{Y}} \sum_{n=1}^{N}\mu_n(\mathbf{y}) + \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} y_{n,t}(m)\mathbb{E}\Bigg[C\Bigg(S_{n,t} - x^*_{n,t}(m) + \sum_{k\neq n}\sum_{j=1}^{M}(S_{k,t} - x^*_{k,t}(j))\mathbb{I}^*_{k,t}(j) + \sum_{k,j} x^*_{k,t+1}(j)\Bigg)\Bigg].$$

Now, that $x^*_{n_0,t_0}(m_0) > 0$, the superposition

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M} y_{n_0,t}(m)\mathbb{E}\Bigg[C\Bigg(S_{n_0,t} - x^*_{n_0,t}(m) + \sum_{k\neq n}\sum_{j=1}^{M}(S_{k,t} - x^*_{k,t}(j))\mathbb{I}^*_{k,t}(j) + \sum_{k,j} x^*_{k,t+1}(j)\Bigg)\Bigg] \quad (23)$$

is dependent on $\{y_{n_0,t_0}(m)\}_m$, thus, yielding the pricing policy $y^*_{n_0,t_0}(m) \neq \hat{y}_{n_0,t_0}(m)$, $\forall m$. But $\mu_n(\mathbf{y})$ is strictly concave in $\mathbf{y}$, and $\hat{\mathbf{y}}$ is its unique maximizer. Thus, $\mu_n(\mathbf{y}^*) < \mu_n(\hat{\mathbf{y}})$.

($\Leftarrow$) If $x^*_{n_0,t}(m) = 0$, $\forall m$, then the superposition (23) is independent of $\{y_{n_0,t}(m)\}$, $\forall m, t$. Hence, $y^*_{n_0,t}(m) = \hat{y}_{n_0,t}(m)$, and $\Delta\mu_n = 0$. ■

Thus, users whose proactive downloads converge to positive values must receive a positive savings gain, whereas users with zero proactive downloads make neither gain nor loss. However, it can be noted that the profit gain reaped by the SP must be met with savings gain reaped by at least one user. This is formalized in the following Theorem.

Theorem 8: Under the proposed multi-user iterative algorithm, $\Delta R > 0$ and $\Delta\mu > 0$ if and only if the set of active users $\mathcal{B}_t(m)$ is non-empty for some time slot and data item.

Proof. In Theorem 5, we have proved the sufficiency part. Here prove the necessity part, that is $\Delta R > 0$ and $\Delta\mu > 0$ imply that $\mathcal{B}_t(m)$ is non-empty for some time slot and data item.

($\Leftarrow$) Suppose that $\mathcal{B}_t(m)$ is empty for all m, t. We have the proposed algorithm $R(\mathbf{x}^{(1)}, \hat{\mathbf{y}}) \geq R(\mathbf{0}, \hat{\mathbf{y}})$, with equality if and only if $\mathbf{x}^{(1)} = \mathbf{0}$. So, suppose towards contradiction that $\mathbf{x}^{(1)} \neq 0$. Hence, that $\eta(\mathbf{0}, \hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)}, \hat{\mathbf{y}})$. By the mean value theorem for random variables, we have $\eta(\mathbf{0}, \hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)}, \hat{\mathbf{y}}) \leq$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Bigg[C'\left(\hat{L}_t\right)\left(\sum_{m=1}^{M}\sum_{n=1}^{N} x^{(1)}_{n,t}(m)\hat{\mathbb{I}}_{n,t}(m) - \sum_{m=1}^{M}\sum_{n=1}^{N} x^{(1)}_{n,t+1}(m)\right)\Bigg] \overset{(a)}{=}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} x^{(1)}_{n,t}(m)(\hat{\mathbb{I}}_{n,t}(m)C'(\hat{L}_t) - C'(\hat{L}_{t-1}))\right]$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} x^{(1)}_{n,t}(m)\mathbb{E}\left[\mathbb{I}_{n,t}(m)C'(\hat{L}_t) - C'(\hat{L}_{t-1})\right] \overset{(b)}{\leq}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\sum_{n\in\mathcal{B}_t(m)}^{N} S_{n,t}\mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m)C'(\hat{L}_t) - C'(\hat{L}_{t-1})\right].$$

Equality (a) follows be rearranging the terms, and inequality (b) follows by summing only over the elements of $\mathcal{B}_t(m)$ which can only increase the RHS. Now, since $\mathcal{B}_t(m)$ is empty for all m, t, it turns out that the RHS of (b) is zero, which is a contradiction. Finally, with $\mathbf{x}^{(1)} = \mathbf{0}$, we have $\mathbf{y}^{(2)} = \hat{\mathbf{y}}$, yielding $\mathbf{x}^{(2)} = \mathbf{0}$, and so on, leading to zeros profit and savings gains. ■

Theorem 8 shows that the existence of active users in the system is necessary and sufficient for profit and savings gains. The attained win-win situation attributes to the judicious utilization of the predictable nature of the user behavior. In particular, the SP trades pricing discounts for certainty which improves proactive download performance, and thus more profit. Users who receive proactive downloads receive also pricing discounts so as to access proactively served content with higher probability. Thus, both parties win.

Remark 4 (Generalizing $\phi_{m,n,t}$): In Theorems 6, 7, 8, we have developed the results given that the proposed algorithm converges to a locally optimal solution of the problem with $\Delta R > 0$, which is attained under the affine form of $\phi_{m,n,t}(y_{n,t}(m))$. However, the same results also hold for any locally optimal solution for the problem with $\Delta R > 0$ and $\phi_{m,n,t}$ taking a more general form of non-negative and non-decreasing function in the employed pricing policy.

Remark 5 (Multiple-slot ahead proactive service): In the analysis above, we have focused on the one-slot ahead proactive service for fresh content guarantees. Yet, for data items with more static content, proactive service can be applied up to ρ slots ahead, $\rho \geq 1$. The profit gain can then be lower bounded through an extended definition of active users in which user n is active with respect to item m and time t if there exists a time slot $t' \in \{t-\rho, \cdots, t\}$ such that $\mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m)C'(\hat{L}_t) - C'(\hat{L}_{t'})\right] > 0$. Thus, this shows more opportunities for existence of active users in the system, and promises of larger scaling orders of the function $g(N)$ of Theorem 6.

Having gained some insights on the performance of the proposed algorithm, and the incentives for users to save money, we consider the distributed implementation of it in the following subsection.

C. Distributed Implementation

The need for distributed implementation of the algorithm is to reduce the centralized complexity at the SP and enable system scalability. Inspired by the single-user scenario where the distributed implementation took place through the sequential solution of price allocation and corresponding optimal proactive downloads, here we also aim to leverage such dynamics to control the algorithm. In particular, we observe that the optimization (17) can be decomposed to price allocation and proactive download control. Hence, we propose to run the distributed version of the multi-user algorithm through the iteration of the two consecutive steps of price and proactive download assignment.

Users will determine the optimal proactive download corresponding to the prices set by the SP, then the SP will respond with optimized pricing. The allocation of optimal proactive download in a distributed fashion is essentially more difficult than in the single-user case, since users need to iterate on the proactive downloads within the proactive download allocation step itself until they converge to the minimizer of the average cost under the current pricing. Thus, there will be an inner loop within the proactive download step.

Also, it can be noted from the previous subsection that not all users may necessarily attain a positive savings gain through the proposed algorithm. Hence, before the proactive download step of each iteration k, the SP can compute the new sets of active users defined as:

$$\mathcal{B}_t^{(k)}(m) := \Big\{ n : \mathbb{E}\left[\mathbb{I}_{n,t}^{(k-1)}(m)C'(L_t^{(k-1)}) - C'(L_{t-1}^{(k-1)})\right] > 0 \Big\}, t = 0, \cdots, T-1, \quad m = 1, \cdots, M, \quad (24)$$

where $L_t^{(k-1)} := \sum_{n=1}^{N} S_{n,t} \sum_{m=1} \mathbb{I}_{n,t}^{(k-1)}(m)$, and $\mathbb{I}_{n,t}^{(k-1)}(m)$ is the indicator that user n requests item m at time t under the pricing policy $\mathbf{y}^{(k-1)}$.

Then, only the active users will be called to participate in such an iteration. The following lemma reveals a nice property of the distributed implementation.

Lemma 4: Starting from the baseline initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) = (\mathbf{0}, \hat{\mathbf{y}})$, the updated sets of active users according to (24) satisfy $\mathcal{B}_t^{(k)}(m) \subseteq \mathcal{B}_t^{(k-1)}(m)$, for all m, t, k.

Proof. By induction. We see from the proof of Theorem 8 that if $n \notin \mathcal{B}_t^{(0)}(m)\ \forall m, t$, i.e., n is not an active user under $\hat{\mathbf{y}}$, then $y_{n,t}^{(1)}(m) = \hat{y}_{n,t}(m)$, $\forall m, t$. Now,

$$\mathbb{E}[\mathbb{I}_{n,t}^{(1)}(m)C'(L_t^{(1)})] - \mathbb{E}[C'(L_{t-1}^{(1)})] = \mathbb{E}[\hat{\mathbb{I}}_{n,t}(m)C'(\hat{L}_t)] - \mathbb{E}[C'(\hat{L}_{t-1})],$$

since $\mathbb{E}[C'(L_{t-1})]$ is independent of the pricing policy. Thus, $n \notin \mathcal{B}_t^{(1)}(m)$, $\forall m, t$. We can also apply the same procedure with all subsequent iterations to show that $n \notin \mathcal{B}_t^{(k)}(m)$, $\forall m, t, k$.

Now, suppose that $\mathcal{B}_t^{(k-1)}(m) \subseteq \mathcal{B}_t^{(k-2)}(m)$, and a pick a user $n \in \mathcal{B}_t^{(k-2)}(m)$ but $n \notin \mathcal{B}_t^{(k-1)}(m)$, for all m, t. Then, clearly, $x_{n,t}^{(k-1)}(m) = 0, \forall m, t$, which in turn implies $\mathbf{y}^{(k)} = \hat{\mathbf{y}}$, and hence $n \notin \mathcal{B}_t^{(k)}(m)$ for the reason that $\mathbb{E}[C'(L_{t-1})]$ is independent of the pricing policy. ■

Thus, users that have been found *not-active* at a given iteration can be safely excluded from participation in the distributed algorithm for the rest of iterations. Consequently, the system will efficiently avoid unnecessary computations. We present our proposed distributed version of the iterative multi-user algorithm in Algorithm 1.

We can see from the algorithm that after each price allocation decision, selected users iterate on optimizing their proactive downloads, with the coordination of the SP, which supplies them with the needed information encapsulated in the personalized cost functions $\eta_n(\mathbf{x}_n, \mathbf{x}_{-n}^{(k,l-1)})$ to minimize, until convergence to the best proactive download policy corresponding to the set prices. Thus users need not exchange their own information with each other. We also note that the iterations on the proactive downloads (i.e., the inner loop) always converge to $\mathbf{x}^{(k)}$, the optimal proactive download of (17) since the problem is convex in $\mathbf{x}$.

Algorithm 1 Distributed Multi-user Iterative Pricing and Proactive Download

1: Initialization: choose the initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, minimum convergence accuracy $\epsilon > 0$, and set the iteration index $k = 1$.
2: **while** $\|(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - (\mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)})\| > \epsilon$ *outer loop* **do**
3: *Service provider:* Compute the sets $\mathcal{B}_t^{(k)}(m)$ as in (24).
4: *Service provider:* Solve

$$\mathbf{y}^{(k)} = \arg\max_{\mathbf{y}} \mu_{\mathbf{y}} - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} h_{n,t}^{(k)}(m) y_{n,t}(m)$$
$$\text{s.t.,} \quad 0 \le y_{n,t}(m) \le v_n(m),$$
$$\frac{1}{M}\sum_{m=1}^{M} y_{n,t}(m) = \bar{y}_{n,t} \forall m, n, t.$$

5: *Service provider:* Set an iteration index $l = 1$ for the inner loop, and let $\mathbf{x}_n^{(k,0)} := (x_{n,t}^{(k)}(m))_{m,t}$, $\forall n \in \cup_{m,t}\mathcal{B}_t^{(k)}(m)$.
6: **while** $\|\mathbf{x}^{(k,l)} - \mathbf{x}^{(k,l-1)}\| > \epsilon$ *inner loop* **do**
7: *Service provider:* To each user $n \in \cup_{m,t}\mathcal{B}_t^{(k)}(m)$ compute $\eta_n(\mathbf{x}_n, \mathbf{x}_{-n}^{(k,l-1)})$.
8: **for** $n \in \cup_{m,t}\mathcal{B}_t^{(k)}(m)$ **do**
9: *User n solve* $\mathbf{x}_n^{(k,l)} := \arg\min_{\mathbf{x}_n} \eta_n(\mathbf{x}_n, \mathbf{x}_{-n}^{(k,l-1)})$

$$\text{s.t.,} \quad x_{n,0}(m) = x_{n,T}(m),$$
$$0 \le x_{n,t}(m) \le S_{n,t}, \quad \forall m, t,$$

where $\mathbf{x}_n = (x_{n,t}(m))_{m,t}$ is the proactive download decision vector for user n, and $\eta_n(\mathbf{x}_n, \mathbf{x}_{-n})$ is the expected SP's cost as a function of $\mathbf{x}_n$, with the proactive downloads for the rest of users $\mathbf{x}_{-n} = (\mathbf{x}_k)_{k \ne n}$ given.
10: **end for**
11: $l = l + 1$.
12: **end while**
13: $k = k + 1$.
14: **end while**

V. Numerical Results

In this subsection, we numerically show the merits of the proposed smart pricing and proactive data downloads through numerical simulations. We first start with simulations for the single user scenario, and then show the gains of the multi-user case.

A. Performance of the single-user case

We consider a setup of $T = 5$ time slots with the user inactivity captured by $(q_t)_t = (0.03, 0.9, 0.02, 0.01, 0.9)$. The number of data items is $M = 3$, the data item size is $S_t = 100$, $\forall t$, the willingness-to-pay values are $(2.99, 2.24, 2.10)$ for the three items, and the average price per item $\bar{y}_t = 2$ for all slots. We consider a quadratic cost function $C(L) = BL^2$ with $B = 2 \times 10^{-4}$.

We run the single-user iterative algorithm starting from the no proactive downloads initial point $(\mathbf{0}, \hat{\mathbf{y}})$. Convergence results of the expected profit as a function are plotted versus the iteration number.

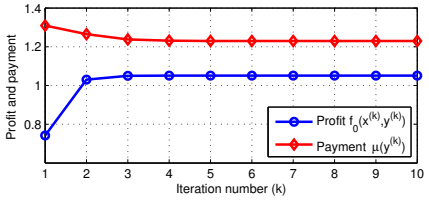

Fig. 1: Convergence of profit and user payment in the single-user scenario.

Fig. 1 depicts the convergence results of the system profit and user payment under the proposed single-user iterative algorithm. Clearly, convergence points imply positive profit gain associated with positive savings gain, i.e., win-win situation.

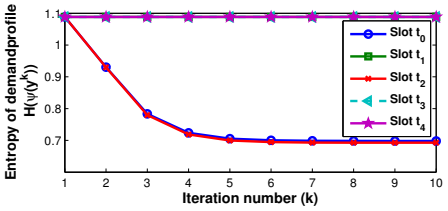

Fig. 2: Entropy of the normalized demand profiles.

To further elaborate on the effect of the pricing strategy on the certainty of the user demand, we evaluate the entropy of the normalized profile $\psi_{m,t} = \phi_{m,t}/(1 - q_t)$ and plot it versus the iteration number [23]. The value of $\psi_{m,t}$ captures the probability of requesting item m at time t conditioned on the event that the user requests at least one item at time t. Fig. 2 plots the entropy of the normalized demand profile for all slots $t = 0, \cdots, 4$ with the iteration number. The peak-hour slots 0 and 2 witness reduced entropy values to improve the proactive download performance. Note that, no proactive downloads are applied for slot 3, since it is preceded by a higher activity slot. The off-peak slots 1 and 4 are not assigned proactive download values too.

B. Performance of the multi-user case

For the multi-user case, we show the convergence results in Figs. 3, 4. In particular, we consider a system with $N = 7$ users and $M = 5$ data items. We assume that users exhibit the same activity probabilities $q_{n,t}$ as in the single-user case, that is, $(q_{n,t})_t = (0.03, 0.9, 0.02, 0.01, 0.9)$, $\forall n$. We also use the quadratic cost function, but with B chosen such that the baseline profit is zero, i.e., $R(\mathbf{0}, \hat{\mathbf{y}}) = 0$.

We have randomly generated the system parameters as

$$[S_{n,t}]_{n,t} = \begin{bmatrix} 8.69 & 9.06 & 39.18 & 53.07 & 0.23 \\ 64.26 & 4.71 & 21.12 & 39.53 & 8.44 \\ 5.81 & 2.16 & 54.11 & 60.39 & 8.10 \\ 9.91 & 4.39 & 13.73 & 22.18 & 0.57 \\ 18.48 & 1.63 & 28.28 & 47.73 & 9.87 \\ 71.12 & 8.56 & 23.13 & 78.42 & 6.89 \\ 1.28 & 8.03 & 45.56 & 41.29 & 9.69 \end{bmatrix},$$

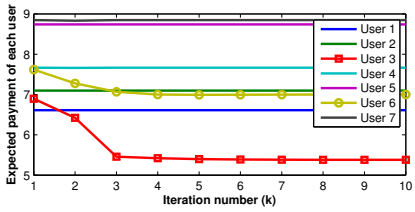

Fig. 3: Convergence of the expected per-user payment.

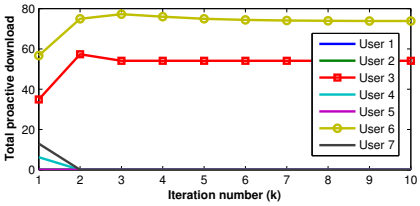

Fig. 4: Convergence of aggregate proactive downloads per user.

$$[v_n(m)]_{n,m} = \begin{bmatrix} 3.40 & 3.59 & 6.16 & 3.1 & 5.25 \\ 2.26 & 6.93 & 3.87 & 2.90 & 5.01 \\ 2.22 & 3.40 & 2.43 & 4.80 & 2.19 \\ 3.24 & 4.69 & 10.43 & 3.30 & 6.03 \\ 5.13 & 2.85 & 12.38 & 2.60 & 10.15 \\ 3.81 & 2.11 & 4.37 & 2.24 & 7.54 \\ 5.31 & 6.37 & 2.04 & 7.39 & 15.17 \end{bmatrix},$$

and $D_{n,t} = D_{n,\tau}$, for all t, τ, with $(D_{n,1})_n = (11.47,\ 10.98,\ 5.04,\ 17.70,\ 23.09,\ 10.01,\ 26.28)$.

In Fig. 3, the evolution of the user payments are plotted throughout the algorithm iterations. It is obvious that users 3 and 6 (which are active according to Definition 1) can achieve strictly positive savings gain, while the rest of the users remain at their original payments. The disparate differences between consumption levels from slot 1 to slot 2 in case of user 3 and from slot 4 to slot 0 in case of user 6 render them perfect candidates to receive proactive downloads. In addition, based on their willingness-to-pay values, the initial normalized entropy of their profiles is least amongst all users, thus SP is more certain about their future demand.

To confirm Theorem 7, we plot the total amount of proactive downloads optimized for each user throughout the iterative algorithm evolution. Clearly, users 3 and 6 who achieved positive savings gains are receiving positive proactive download service.

We also have considered the scaling of system gains with the number of users. We run the simulation for $T = 4$ slots, with $q_{n,t} = q_t$, $\forall n$, and $(q_t)_t = (0.32, 0.9, 0.01, 0.95)$, and we consider $M = 4$ items. We take $B = 10^{-3}$, and generate $v_n(m)$ as follows $v_n(m) = 2 + e_n(m)$, where $e_n(m)$ is a realization of an exponential distribution with mean m. We also set $D_{n,t} = \sum_{m=1}^{M} e_n(m)$, $\forall t$. The amount of consumption for user n and time t is generated based on $1-q_t$ as follows. If $q_t < 0.5$ then $S_{n,t}$ is uniformly distributed on $[50, 70]$, and $S_{n,t}$ is uniformly distributed on $[0, 10]$ if $q_t \geq 0.5$. The idea behind such a choice is to increase the likelihood of active users scaling linearly with N. We averaged the results over a multitude of simulation runs and plotted the profit and savings gains in Figs. 5, 6. Further, we also plot the

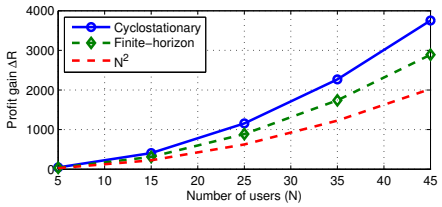

Fig. 5: Expected profit gain grows with the number of users as N^2

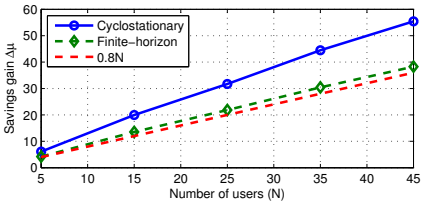

Fig. 6: Expected savings gains vs. the number of users N

results of the *finite-horizon* scenario, when cyclostationarity of demand profiles does not hold possibly because of fast varying willingness-to-pay values, thus the SP needs to optimize each cycle of T-slots independently from the previous one.

We can see from Fig. 5 that the profit gain satisfies the established lower bound in Theorem 6 as it grows at least as N^2, where N is the number of users. Scaling of savings gain is shown in Fig. 6 with an illustrating linear curve of $0.8N$ to give an insight on the growth rate of the aggregate savings. Obviously, there is a linear increase of such gains, which essentially promises of constant saving shares per user. In addition, finite-horizon optimization attains considerable gains despite the limitation of independent resource allocation over cycles. In this result, the average savings gain per user are 30.2% for cyclostationary case, and 20.6% for finite-horizon case. Further, the percentage of average exccess load due to proactive download inaccuracies are 9.4% and 3.4% for the two respective cases. That is, the battery loss associated with proactive download is proportional to $\sim 9.4\%$, $\sim 3.4\%$ of additional load consumption.

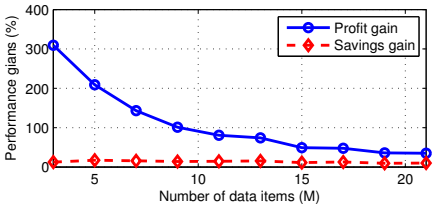

Fig. 7: Impact of number of data items M on performance gains.

In Fig. 7, we study the impact of number of data items M when $N = 12$ users. As M grows, SP suffers more uncertainty and loses profit gain. However, savings gain is not significantly impacted since SP still needs to maintain a balance between high pricing incentives, to combat against uncertainty, and service costs.

VI. Conclusion

In this work, we have addressed the profit maximization problem for a SP that experiences time varying, yet predictable, demand which assumes high peak-to-average ratio. The SP has been allowed to employ joint smart pricing and proactive downloads so as to balance its load level over time, without enforcing users to physically change their regular demand activities. To achieve enhanced performance of proactive caching, the SP assigns personalized pricing policies based on the user preferences of different types of content so as to enhance the certainty about future demand, and concurrently maximize its profit. We have analyzed such a system under single- and multi-user setups and developed efficient algorithms that yield strictly improved profit gain for the SP and savings gain for the end user as compared with the baseline scenario of no-proactive services. We have established bounds on the asymptotic scaling performance of the profit gain, and showed that all users that receive proactive data services must attain positive savings gain.

Appendix A
Proof of Theorem 4

First, we note that $\tilde{R}^{(k)}$ is concave in $(\mathbf{x},\mathbf{y})$ since $\mu(\mathbf{y})$ is a quadratic form of $\mathbf{y}$ with a negative definite Hessian, and $-\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m,n} h_{n,t}^{(k)}(m)y_n,t(m)$ is an affine function of $\mathbf{y}$. Finally, $\eta(\mathbf{x},\mathbf{y}^{(k-1)})$ is a convex function of $\mathbf{x}$, by the definition of the cost function C.

Second, we consider the three conditions specified in Lemma 2. Since R is continuous in $(\mathbf{x},\mathbf{y})$ and is defined over a bounded and closed feasible set, then it has a global maximum value $U>0$. Such a value can be subtracted from $\tilde{R}^{(k)}$ defined above to keep Condition 1) of Lemma 2 satisfied. However, subtracting a constant from the objective function does not affect the solution, which is the main point of interest. Therefore, Condition 1) of Lemma 2 is not necessary in this case. For Condition 2) of Lemma 2, we have

$$\left.\frac{\partial\tilde{R}^{(k)}(\mathbf{x},\mathbf{y})}{\partial x_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})} = \left.\frac{\partial\eta(\mathbf{x},\mathbf{y}^{(k-1)})}{\partial x_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})} = \left.\frac{\partial R(\mathbf{x},\mathbf{y})}{\partial x_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})}, \quad \forall m,n,t.$$

Likewise,

$$\left.\frac{\partial\tilde{R}^{(k)}(\mathbf{x},\mathbf{y})}{\partial y_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})} = \left.\frac{\partial\mu(\mathbf{y})}{\partial y_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})} - \frac{\partial\eta(\mathbf{x}^{(k-1)},\mathbf{y})}{\partial y_{n,t}(m)} = \left.\frac{\partial R(\mathbf{x},\mathbf{y})}{\partial y_{n,t}(m)}\right|_{(\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)})}, \quad \forall m,n,t.$$

Thus Condition 2) of Lemma 2 is satisfied.

Finally, Condition 3) of the same lemma need not be satisfied since it is mainly stated in Theorem 1 [24] for a non-convex constraint function that has to be replaced by a convex approximate. Condition 3) mainly implies the satisfaction of the complementary slackness conditions by both the approximate and the original constraint functions. Since we are interested only in the objective function, Condition 3) of Lemma 2 is not necessary for convergence to a KKT point.

Appendix B
Proof of Theorem 5

We have $\Delta R \geq R(\mathbf{x}^{(1)},\hat{\mathbf{y}}) - R(\mathbf{0},\hat{\mathbf{y}})$, where $\mathbf{x}^{(1)}$ is the proactive download solution to the first iteration of the proposed multiuser algorithm when the starting point is $(\mathbf{0},\hat{\mathbf{y}})$. Thus, $\mathbf{x}^{(1)}$ is also the minimizer of $\eta(\mathbf{x},\hat{\mathbf{y}})$ under the constraints of (17).

Hence, we have $\Delta R \geq \eta(\mathbf{0},\hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)},\hat{\mathbf{x}})$. Further, suppose we use the suboptimal proactive download policy $\tilde{\mathbf{x}}$ rather than $\mathbf{x}^{(1)}$ at the first iteration, whereby $\tilde{x}_{n,t}(m) := \tilde{x}_t, n\in\mathcal{B}_t(m)$, and $\tilde{x}_{n,t}(m)=0$ otherwise. Hence, the profit gain satisfies $\Delta R \geq$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Bigg[C(\hat{L}_t) - C\Bigg(\hat{L}_t - \sum_{m=1}^{M}\sum_{n\in\mathcal{B}_t(m)}\tilde{x}_t\hat{\mathbb{I}}_{n,t}(m) + \sum_{m=1}^{M}\sum_{n\in\mathcal{B}_{t+1}(m)}\tilde{x}_{t+1}\Bigg)\Bigg] \overset{(a)}{\geq}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Bigg[C'\Bigg(\hat{L}_t - \sum_{m=1}^{M}\sum_{n\in\mathcal{B}_t(m)}\tilde{x}_t\hat{\mathbb{I}}_{n,t}(m) + \sum_{m=1}^{M}\sum_{n\in\mathcal{B}_{t+1}(m)}\tilde{x}_{t+1}\Bigg)\times \Bigg(\sum_{m=1}^{M}\sum_{n\in\mathcal{B}_t(m)}\tilde{x}_t\hat{\mathbb{I}}_{n,t}(m) - \sum_{m=1}^{M}\sum_{n\in\mathcal{B}_{t+1}(m)}\tilde{x}_{t+1}\Bigg)\Bigg] \overset{(b)}{\geq}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\tilde{x}_t \sum_{n\in\mathcal{B}_t(m)}\mathbb{E}\Bigg[\hat{\mathbb{I}}_{n,t}(m)C'\Bigg(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(k)}\tilde{x}_t\hat{\mathbb{I}}_{k,t}(j)\Bigg) - C'\Bigg(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)}\tilde{x}_t\Bigg)\Bigg] > 0.$$

Inequality (a) holds by the first order condition on the convexity of the cost function C. Inequality (b) follows by rearranging the terms of the RHS of Inequality (a) and replacing the terms $\sum_{j=1}^{M}\sum_{k\in\mathcal{B}_{t+1}(j)}\tilde{x}_{t+1}$, $-\sum_{j=1}^{M}\sum_{k\in\mathcal{B}_{t-1}(k)}\tilde{x}_{t-1}\mathbb{I}_{k,t-1}(j)$ with zeros while noting that C' is monotonically increasing on its domain. Finally, the last strict inequality holds by the definition of $\tilde{x}_t$ in Theorem 5, and the definition of active users.

Appendix C
Proof of Theorem 6

We have $\Delta R \geq R(\mathbf{x}^{(1)},\hat{\mathbf{y}}) - R(\mathbf{0},\hat{\mathbf{y}})$, this reduces to $\Delta R(N) \geq \eta(\mathbf{0},\hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)},\hat{\mathbf{y}})$, the cost reduction after the first iteration of the multi-user iterative algorithm. To show that $\eta(\mathbf{0},\hat{\mathbf{y}}) - \eta(\mathbf{x}^{(1)},\hat{\mathbf{y}}) = \Omega(g(N)C'(\gamma N))$, we consider two steps. In the first step, we show that if $\liminf_{N\to\infty}\tilde{x}_t > 0$

then

$$\liminf_{N\to\infty} \frac{1}{g(N)C'(\gamma\cdot N)} \sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t\hat{\mathbb{I}}_{k,t}(j)\right)\right] > 0,$$

for some $\gamma > 0$. In the other step, we prove that $\liminf_{N\to\infty}\tilde{x}_t > 0$, and

$$\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t\hat{\mathbb{I}}_{k,t}(j)\right)\right]}{\sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[C'\left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t\right)\right]} > 1.$$

Step 1: Suppose that $\liminf_{N\to\infty}\tilde{x}_t > 0$. By Fubini's theorem, we can move the summation inside the expectation, since all the summands are non-negative. Also, by Fatou's lemma, we have

$$\begin{aligned}
&\liminf_{N\to\infty} \frac{1}{g(N)C'(\gamma\cdot N)} \\
&\sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \tilde{x}_t\hat{\mathbb{I}}_{k,t}(j)\right)\right] \geq \\
&\mathbb{E}\left[\liminf_{N\to\infty} \frac{\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)}{g(N)} \times \right. \\
&\left.\liminf_{N\to\infty} \frac{C'\left(\sum_{j=1}^{M}\sum_{k=1,k\neq n}^{N}(S_{k,t}-\tilde{x}_t)\hat{\mathbb{I}}_{k,t}(j)\right)}{C'(\gamma N)}\right] \overset{(a)}{=} \\
&\liminf_{N\to\infty} \frac{\sum_{n\in\mathcal{B}_t(m)} P_{n,t}(m)}{g(N)} \times \\
&\mathbb{E}\left[\liminf_{N\to\infty} \frac{C'\left(N\cdot\sum_{j=1}^{M}\frac{\sum_{k=1,k\neq n}^{N}(S_{k,t}-\tilde{x}_t)\hat{\mathbb{I}}_{k,t}(j)}{N}\right)}{C'(\gamma N)}\right].
\end{aligned}$$

Note that, on the left hand side (LHS) of Equality (a), we have removed the contribution of user n in the argument of C' which can only reduce its value, thus yielding $\hat{\mathbb{I}}_{n,t}(m)$ independent of the argument of C'. Hence, on the RHS, we split the expectation over the product. But $1-q_{n,t} > \epsilon$ for all n,t, for any $n\in\mathcal{B}_t(m)$, $\liminf_{N\to\infty}\check{P}(m) > 0$, and $\beta_t(m) > 0$ by hypothesis. Therefore,

$$\liminf_{N\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{P_{n,t}(m)}{g(N)} = \beta_t(m) \liminf_{B_t(m)\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{P_{n,t}(m)}{B_t(m)} > 0.$$

On the other hand, Kolmogorov's strong law of large numbers implies

$$\begin{aligned}
\gamma :&= \lim_{N\to\infty} \sum_{j=1}^{M} \frac{\sum_{k=1,k\neq n}^{N}(S_{k,t}-\tilde{x}_t)\hat{\mathbb{I}}_{k,t}(j)}{N} \\
&\overset{a.s.}{=} \lim_{N\to\infty} \sum_{j=1}^{M} (S_{k,t}-\tilde{x}_t)\frac{\sum_{k\neq n} P_{k,t}(j)}{N} > 0,
\end{aligned}$$

since $q_{n,t} < 1, \forall n,t$. Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[\liminf_{N\to\infty} \frac{\sum_{n\in\mathcal{B}_t(m)} P_{n,t}(m)}{g(N)} \times \right. \\
&\left.\liminf_{N\to\infty} \frac{C'\left(N\cdot\sum_{j=1}^{M}\frac{\sum_{k=1,k\neq n}^{N}(S_{k,t}-\tilde{x}_t)\hat{\mathbb{I}}_{k,t}(j)}{N}\right)}{C'(\gamma N)}\right] \geq \\
&\beta_t(m)\cdot\liminf_{B_t(m)\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{P_{n,t}(m)}{B_t(m)} > 0.
\end{aligned}$$

Step 2: In this step, we prove that there exists $\chi > 0$, independent of N, for which if $\tilde{x}_t = \chi$, then

$$\liminf_{N\to\infty} \frac{\sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\hat{\mathbb{I}}_{n,t}(j)\right)\right]}{\sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[C'\left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\right)\right]} > 1.$$

We set $\tilde{x}_t = \chi$, independent of N, and we will prove that $0 < \chi < \check{S}$. By Fubini's theorem and Fatou's lemma, as in ***Step 1***, it suffices to prove that

$$\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\hat{\mathbb{I}}_{n,t}(j)\right)\right]}{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} C'\left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\right)\right]} > 1,$$

for some $0 < \chi < \check{S}$.

By Condition (21), we have

$$\begin{aligned}
&\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)C'\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\hat{\mathbb{I}}_{k,t}(j)\right)\right]}{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} C'\left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\right)\right]} = \\
&\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)\left(\hat{L}_t - \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\hat{\mathbb{I}}_{k,t}(j)\right)^{\delta}\right]}{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)} \chi\right)^{\delta}\right]} \overset{(a)}{\geq} \\
&\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \hat{\mathbb{I}}_{n,t}(m)\left(\sum_{j=1}^{M}\sum_{k=1}^{N}(S_{k,t}-\chi)\hat{\mathbb{I}}_{k,t}(j)\right)^{\delta}\right]}{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \left(\hat{L}_{t-1} + \sum_{j=1}^{M}\sum_{k=1}^{N} \chi\right)^{\delta}\right]} \overset{(b)}{=} \\
&\liminf_{N\to\infty} \frac{\mathbb{E}\left[\sum_{n\in\mathcal{B}_t(m)} \frac{\hat{\mathbb{I}}_{n,t}(m)}{B_t(m)}\left(\sum_{j=1}^{M}\sum_{k=1}^{N}\frac{(S_{k,t}-\chi)\hat{\mathbb{I}}_{k,t}(j)}{N}\right)^{\delta}\right]}{\mathbb{E}\left[\left(\sum_{j=1}^{M}\sum_{k=1}^{N}\frac{S_{k,t-1}\hat{\mathbb{I}}_{k,t-1}(j)}{N} + M\chi\right)^{\delta}\right]},
\end{aligned}$$

where Inequality (a) follows since we extend the negative sum $-\sum_{j=1}^{M}\sum_{k\in\mathcal{B}_t(j)}\chi\hat{\mathbb{I}}_{n,t}(j)$ to include the terms from

outside set $\mathcal{B}_t(j)$, $j = 1, \cdots, M$, while noting that C' is monotonically increasing. Further, we increase denominator by extending the sum to $\sum_{j,k} \chi$. Equality (b) follows through multiplying both the numerator and denominator by $N^\delta B_t(m)$.

Now, by Kolmogorov's strong law of large numbers, we define the quantities:

$$c_1(m) := \liminf_{N\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{\hat{\mathbb{I}}_{n,t}(m)}{B_t(m)} \overset{a.s.}{=} \liminf_{N\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{P_{n,t}(m)}{B_t(m)},$$

$$c_2(j) := \limsup_{N\to\infty} \sum_{k=1}^{N} \frac{\hat{\mathbb{I}}_{k,t-1}(j)}{N} \overset{a.s.}{=} \limsup_{N\to\infty} \sum_{k=1}^{N} \frac{P_{k,t-1}(j)}{N},$$

$$c_3(j) := \liminf_{N\to\infty} \sum_{k=1}^{N} \frac{\hat{\mathbb{I}}_{k,t}(j)}{N} \overset{a.s.}{=} \liminf_{N\to\infty} \sum_{k=1}^{N} \frac{P_{k,t}(j)}{N}.$$

Since $1-q_{n,\tau} < \epsilon$, $\forall n, \tau$, $\beta_t(m) > 0$, and $\liminf_{N\to\infty} \check{P}_t(m) \overset{a.s.}{>} \epsilon$, $\forall k \in \mathcal{B}_t(m)$, we conclude $c_1(m) \overset{a.s.}{>} 0$, $c_2(j) > 0$, $j = 1, \cdots, M$, and $c_3(j) \overset{a.s.}{>} 0 j = 1, \cdots, M$. Hence,

$$\frac{\mathbb{E}\left[\liminf_{N\to\infty} \sum_{n\in\mathcal{B}_t(m)} \frac{\hat{\mathbb{I}}_{n,t}(m)}{B_t(m)} \left(\sum_{j=1}^{M}\sum_{k=1}^{N} \frac{(S_{k,t}-\chi)\hat{\mathbb{I}}_{k,t}(j)}{N}\right)^\delta\right]}{\mathbb{E}\left[\limsup_{N\to\infty} \left(\sum_{j=1}^{M}\sum_{k=1}^{N} \frac{S_{k,t-1}\hat{\mathbb{I}}_{k,t-1}(j)}{N} + M\chi\right)^\delta\right]} \overset{(c)}{=} \frac{c_1(m)\left(\sum_{j=1}^{M}(S_3-\chi)c_3(j)\right)^\delta}{\left(M\chi + \sum_{j=1}^{M} S_2 c_2(j)\right)^\delta},$$

where $S_3 := \liminf_{N\to\infty} \sum_{k=1}^{N} S_{k,t}/N$, and $S_2 := \limsup_{N\to\infty} \sum_{k=1}^{N} S_{k,t-1}/N$. The right hand side (RHS) of Equality (c) is strictly greater than 1 if and only if

$$\chi < \frac{\left(c_1(m)^{\frac{1}{\delta}} \sum_{j=1}^{M} c_3(j) S_3\right) - \sum_{j=1}^{M} c_2(j)) S_2}{\left(M + \sum_{j=1}^{M} c_3(j)\right)}.$$

Now, to show that $\chi > 0$, it suffices to prove that

$$\left(c_1(m)^{\frac{1}{\delta}} \sum_{j=1}^{M} c_3(j) S_3\right) - \sum_{j=1}^{M} c_2(j) S_2 > 0.$$

We have by the definition of set $\mathcal{B}_t(m)$ that

$$\sum_{n\in\mathcal{B}_t(m)} \mathbb{E}\left[\hat{\mathbb{I}}_{n,t}(m) C'(\hat{L}_t) - C'(\hat{L}_{t-1})\right] > 0.$$

By Condition (21), consider $n \in \mathcal{B}_t(m)$, for sufficiently large N, we obtain

$$P_{n,t}(m) > \mathbb{E}\left[\left(\hat{L}_{t-1}\right)^\delta\right] / \mathbb{E}\left[\left(\hat{S} + \hat{L}_t\right)^\delta\right]$$

$$\text{which implies } \sum_{n\in\mathcal{B}} \frac{P_{n,t}(m)}{B_t(m)} > \frac{\mathbb{E}\left[\left(\hat{L}_{t-1}\right)^\delta\right]}{\mathbb{E}\left[\left(\hat{S} + \hat{L}_t\right)^\delta\right]}$$

$$\liminf_{N\to\infty} \sum_{n\in\mathcal{B}} \frac{P_{n,t}(m)}{B_t(m)} > \left(\frac{\left(\sum_{j=1}^{M} c_2(j) S_2\right)}{\left(\sum_{j=1}^{M} c_3(j) S_3\right)}\right)^\delta,$$

by (22). Hence, $(c_1(m))^{\frac{1}{\delta}} \frac{\left(\sum_{j=1}^{M} c_3(j) S(j)\right)}{\left(\sum_{j=1}^{M} c_2(j) S(j)\right)} > 1$.

REFERENCES

[1] Cisco Visual Networking Index: Forecast and Methodology, 2012-2017.

[2] D. Niyato, and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of Nash equilibrium, and collusion," *IEEE Journal onSelected Areas in Communications,* vol.26, no.1, pp. 192–202, Jan. 2008

[3] I.C. Paschalidis, and J. N. Tsitsiklis, "Congestion-dependent pricing of network services," *IEEE/ACM Transactions on Networking,* vol. 8, no. 2, pp. 171–184, Apr 2000.

[4] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time Dependent Pricing for Mobile Data," *ACM SIGCOMM 2012*.

[5] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 357–362, 14-19 April 2013.

[6] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 351–356, 14-19 April 2013.

[7] G. Pallis, and A. Vakali, Insight and perspectives for content delivery networks, *Communications of the ACM*, vol. 49, vo. 1, ACM Press, NY, USA, pp. 101-106, January 2006.

[8] A. Vakali, and G. Pallis, Content delivery networks: status and trends, *IEEE Internet Computing*, IEEE Computer Society, pp. 68-74, November-December 2003.

[9] A. K. Pathan, and R. Buyya, A taxonomy and survey of content delivery networks, *Tech Report*, University of Melbourne, 2007.

[10] J. Kangasharju, J. Roberts, and K. Ross, "Object replication strategies in content distribution networks," *Computer Communications*, vol 25, no. 4, pp. 376383, March 2002.

[11] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive content distribution for dynamic content", *2013 IEEE International Symposium on Information Theory (ISIT)*, pp. 1232–1236, July 2013.

[12] Y. Bao, X. Wang, S. Zhou, and Zhisheng Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," *2012 18th Asia-Pacific Conference on Communications (APCC)*, vol., no., pp.566,571, 15-17 Oct. 2012.

[13] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive resource allocation: harnessing the diversity and multicast gains," *IEEE Transactions on Information Theory,* vol. 59, no. 8, pp. 4833–4854, Aug. 2013.

[14] J. Tadrous, A. Eryilmaz and H. El Gamal, "Proactive Content Download and User Demand Shaping for Data Networks," to appear, *IEEE/ACM Transactions on Networking*.

[15] Longbo Huang, Shaoquan Zhang, Minghua Chen, and Xin Liu, "When backpressure meets predictive scheduling," *In Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc '14)*, ACM, New York, NY, 2014.

[16] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Pricing for demand shaping and proactive download in smart data networks," *Proceedings IEEE INFOCOM 2013*, pp. 3189–3194, April 2013.

[17] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Joint pricing and proactive caching for data services: global and user-centric approaches," *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 616–621, May 2014.

[18] http://oss.oetiker.ch/rrdtool/gallery/index.en.html

[19] C. Song, Z. Qu, N. Blumm, and A. Barabas, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018-1021, Feb. 2010.

[20] S. Boyd and L. Vandenberge, "Convex Optimization," *Cambridge University Press*, 2004.

[21] A. Di Crescenzo,"A probabilistic analogue of the mean value theorem and its applications to reliability theory," *Journal of Applied Probability*, vol. 36, no. 3, pp. 706–719, Sept. 1999.

[22] A. MacKenzie, and L. DaSilva, "Game theory for wireless egnineers," *Morgan & Claypool*, 2006.

[23] T. Cover and J. Thomas, "Elements of information theory," *Wiley − Interscience*, 2006.

[24] B. Marks and G. Wright, "A general inner approximation algorithm for nonconvex mathematical programs, *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

[25] C. Breidert, M. Hahsler and T. Reutterer, "A review of methods for measuring willingness-to-pay," *Innovative Marketing*, vol. 2, no. 4, pp. 8–32, 2006.

[26] G. Yan-yan and L. Qi-cheng, "E-commerce personalized recommendation system based on multi-agent", *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, vol. 4, pp. 1999–2003, Aug. 2010.
[27] http://www.netflixprize.com/

John Tadrous (S'10– M'15) has received his PhD degree in electrical engineering from the ECE Department at The Ohio State University in 2014, MSc degree in wireless communications from the Center of Information Technology at Nile University in 2010, and BSc degree from the EE Department at Cairo University. He was a research assistant at the Wireless Intelligent Networks Center (WINC) at Nile University between 2008 and 2010, where he worked on resource allocation and power control for cognitive radio networks. Between 2010 and 2014 he was a research associate at the Information Processing Systems Lab, where he worked on proactive resource allocation and scheduling, smart data pricing, and information theory. Since May 2014 he has joined the Center for Multimedia Communication (CMC) at Rice University as a post-doctoral research associate, where he works on modeling and analysis of interactive data traffic, full-duplex communications, and downlink channel estimation for massive MIMO systems.

Atilla Eryilmaz (S'00–M'06) received his B.S. degree in Electrical and Electronics Engineering from Boğaziçi University, Istanbul, in 1999, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2001 and 2005, respectively. Between 2005 and 2007, he worked as a Postdoctoral Associate at the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology. He is currently an Associate Professor of Electrical and Computer Engineering at The Ohio State University, where he has been a faculty since 2007. He served as a TPC chair for the International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt) 2015, and is an Associate Editor for ACM/IEEE Transactions on Networking since March 2015. Dr. Eryilmazs research interests include design and analysis for complex networked systems with focus on wireless communication and power networks, optimal control of stochastic networks, optimization theory, distributed algorithms, network pricing, and information theory. He was a co-author of the Best Student Paper Award in WiOpt 2012. He received the NSF-CAREER Award in 2010, and two Lumley Research Awards for Research Achievements in 2010 and 2015.

Hesham El Gamal (M99–SM03–F10) received the B.S. and M.S. degrees in electrical engineeringfrom Cairo University, Cairo, Egypt, in 1993 and 1996, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Maryland at College Park, MD, in 1999. From 1993 to 1996, he served as a Project Manager in the Middle East Regional Office of Alcatel Telecom. From 1996 to 1999, he was a Research Assistant in the Department of Electrical and Computer Engineering, the University of Maryland at College Park, MD. From February 1999 to December 2000, he was with the Advanced Development Group, Hughes Network Systems (HNS), Germantown, MD, as a Senior Member of Technical Staff. Since January 2001, he has been with the Electrical and Computer Engineering Department at Ohio State University where he is now a Professor. He held visiting appointments at UCLA, Institut Eurecom, and served as a Founding Director for the Wireless Intelligent Networks Center (WINC) at Nile University (2007-2009). Dr. El Gamal is a recipient of the HNS Annual Achievement Award (2000), the OSU College of Engineering Lumley Research Award (2003, 2008), the OSU Electrical Engineering Department FARMER Young Faculty Development Fund (2003-2008), the OSU Stanley E. Harrison Award (2008), and the National Science Foundation CAREER Award (2004). He holds 12 patents and has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS (2001-2005), an Associate Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING (2003-2007), a Guest Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY SPECIAL ISSUE ON COOPERATIVE COMMUNICATIONS (2007), a member of the SP4COM technical committee (2002–2005), a cochair of the Globecom08 Communication Theory Symposium, and a cochair of the 2010 IEEE Information Theory Workshop. He served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY.