On Optimal Proactive Caching for Mobile Networks With Demand Uncertainties

John Tadrous, *Member, IEEE*, and Atilla Eryilmaz, *Member, IEEE*

***Abstract*—Mobile data users are known to possess predictable characteristics both in their interests and activity patterns. Yet, their service is predominantly performed, especially at the wireless edges, "reactively" at the time of request, typically when the network is under heavy traffic load. This strategy incurs excessive costs to the service providers to sustain *on-time* (or *delay-intolerant*) delivery of data content, while their resources are left underutilized during the light-loaded hours. This motivates us in this work to study the problem of optimal "proactive" caching whereby, future delay-intolerant data demands can be served within a given prediction window *ahead of* their actual time-of-arrival to minimize service costs. To that end, we first establish fundamental bounds on the minimum possible cost achievable by any proactive policy, as a function of the *prediction uncertainties*. These bounds provide interesting insights on the impact of *uncertainty* on the maximum achievable proactive gains. We then propose specific proactive caching strategies, both for uniform and fluctuating demand patterns, that are asymptotically-optimal in the limit as the prediction window size grows while the prediction uncertainties remain fixed. We further establish the exponential convergence rate characteristics of our proposed solutions to the optimal, revealing close-to-optimal performance characteristics of our designs even with small prediction windows. Also, proactive design is contrasted with its reactive and delay-tolerant counter-parts to obtain interesting results on the unavoidable costs of uncertainty and the potentially remarkable gains of proactive operation.**

***Index Terms*—predictable demand, proactive caching, resource allocation, scheduling, uncertainty.**

I. Introduction

RECENTLY, the wireless spectrum has been witnessing tremendous demand to support the emerging throughput-hungry applications (e.g., HD video streaming), which are dominating the wireless data traffic nowadays. By 2016, traffic from wireless and mobile devices is projected to exceed that of wired devices, and the demand on wireless data traffic is expected to multiply by 13-fold between 2012 and 2017 [1]. Coupling these findings with the fact that the available spectrum for wireless communications is a limited resource, a major spectrum shortage problem is facing the wireless communication industry.

It is well documented by the FCC that the wireless spectrum consistently incurs periods of underutilization on a daily basis [2], [3], which is attributed to the users' behavioral patterns as most users idle together in the off-peak times. This study is also strengthened by the wireless spectrum measurements carried out by RRDTool [4], [5], the wireless spectrum tracking client, and the recent data traces collected by major European operators in [6]. Thus, the demand on wireless spectrum varies between a peak level at which service providers incur excessive costs to provide reliable delivery of data content, and an off-peak level at which the precious resource is left underutilized.

There has been extensive research to tackle such a problem, some of which has particularly considered offering pricing incentives for end-users to shift their demand to the off-peak times. Of these, the cognitive radio approach [7]–[9] enables out-of-band users to enhance the utilization of the spectrum in the off-peak times through low-priced service. Attempts as in [10] and [11] jointly assign pricing and scheduling of data services to flatten the demand fluctuations over time. In particular, pricing incentives are traded for extra delay tolerance, hence scheduling policies can be optimized over longer time horizon and consequently attain reduced cost performance.

WiFi offloading [12]–[15] has also gained considerable attention to mitigate the contention on the limited spectrum of wireless carriers in the peak hour, and ideas about rescheduling of carriers' traffic through WiFi networks have been studied. However, WiFi coverage is not present in several outdoor locations where impact of peak hour traffic congestions is severe. For instance, public transit riders suffer degraded QoS since all their wireless access has to be routed through the cellular network. In addition, WiFi networks suffer the same large peak to off-peak demand ratio that requires particular attention to the temporal aspect of content service.

In the aforementioned approaches, scheduling of wireless demand is applied *reactively* so that data requests are initiated beforehand, then the service provider utilizes the leveraged delay tolerance from end-users to schedule them efficiently. Numerous back pressure and virtual queuing techniques have been developed to tackle a variety of network optimization objectives (cf. [16], [17], and the references therein). Thus, cost reduction comes at the expense of disturbed user activity patterns as the service is postponed to off-peak times, or the next available WiFi connection [12].

Despite the predominance of reactive solutions in mobile data services, it is also well known that data users possess

Manuscript received November 28, 2014; revised May 08, 2015; July 13, 2015; and September 08, 2015; accepted September 09, 2015; approved by IEEE/ACM Transactions on Networking Editor X. Wang. Date of publication September 29, 2015; date of current version October 13, 2016. This work was supported by the NSF under Grants CAREER-CNS-0953515, CNS-WiFiUS-1456806, and CCSS-EARS-1444026; the QNRF under Grants NPRP 09-1168-2-455 and NPRP 7-923-2-344; and the DTRA under Grant HDTRA1-15-1-0003.

J. Tadrous is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77025 USA (e-mail: jtadrous@rice.edu).

A. Eryilmaz is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: eryilmaz.2@osu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2015.2478476

consistent (therefore statistically predictable) interests and activity patterns (e.g., [6] and [18]–[20]). This has motivated a few recent works [21], [22] to develop *proactive* scheduling strategies to smooth out the network traffic over time, reduce the service costs, and essentially preserve the users' activity patterns undisturbed. Such an approach has also been implemented practically by some content provider [23], [24].

In the proactive operation paradigm, rather than reactively responding to incoming demands or postponing services, the service provider utilizes the statistically predictable, albeit uncertain, nature of future user demands to prefetch the predictable demand of end-users during off-peak times so that it can be served in part from the local memory upon request. Hence, end-users need not change their regular demand activities.

Earlier works in this new domain, however, have focused on the scenario of perfectly predictable demand where service providers possess full certainty about the future demand instants of each user through a predetermined prediction window size. Yet, prediction uncertainties not only exist almost unavoidably, but they also fundamentally change the nature of the problem as they raise *the possibility of wasting resources by proactively serving undesired data*.

Recent works such as [25] and [26] have considered the impact of uncertainty about the exact user demand where service providers offer valuations and pricing incentives to enhance predictability of user requests over a set of data items. Yet, uncertainty about user activity, that is, *whether the user will request content at all or not*, is not well studied. Furthermore, developed algorithms in those works have been limited to offline (static) implementation with one slot-ahead proactive service.

In this work, we give particular attention to the impact of uncertainty about user activity and consider the design of online (dynamic) proactive strategies that can optimally balance the gains of low-cost transmissions with the risk of unnecessary resource consumption due to prediction uncertainties. In particular, we study the unavoidable costs of uncertainty due to imperfect prediction of user activity, even when service providers manage to achieve perfect knowledge about the content to be consumed. Our model also generalizes the proactive download window size to more than one slot ahead.

In particular, we consider the generic scenario (described in Section II) of a service provider that provides "*delay-intolerant*" (also called *on-time*) services[1] to a group of users who generate possibly time-varying requests that are predictable T time-slots ahead of time, but with uncertainties. The main objective (also in Section II) is for the service provider to perform proactive service decisions depending on the degree of uncertainty about future requests to minimize its expected convex cost over time while maintaining on-time delivery of requested content.

In Section III, we address the basic prediction scenario in which each user demand arrives *uniformly* over time so that we can isolate the impact of prediction uncertainty on the design and performance. For that model, we establish a global lower bound that captures the impact of demand uncertainties on the optimal attainable performance. Moreover, we develop an asymptotically optimal stationary policy that achieves the lower bound as T grows to infinity with an exponential convergence speed. Furthermore, we contrast the performance of our design with its reactive and the infinitely delay-tolerant counterparts to reveal the impact of uncertainty on proactive gains.

Then, in Section IV, we extend the previous scenario to the more realistic case of *fluctuating* demand patterns in order to explore the impact of peak and off-peak differences on the proactive design and performance. We establish a global lower bound for this prediction model and show that it has a significant potential to minimize the cost below that of Section III. We develop an asymptotically optimal cyclostationary proactive caching policy that attains such lower bound. Similar to the uniform demand case, we also establish the convergence speed of the developed policies to the lower bound to be exponential in T.

II. System Model and Problem Formulation

We consider the time-slotted operation of a network comprising a service provider, that provides data content to end-users, e.g., YouTube, Netflix, CNN, Facebook, ESPN, etc., and a set $\mathcal{N} = \{1, \ldots, N\}$ of N users. In our design, we focus primarily on large timescale of operation whereby the time-slot duration is comparable to the duration of content consumption, which may range from minutes to hours depending on the nature of the services.

As we consider data content providers, remain in system for much larger durations than that of a time-slot. In particular, users subscribe through monthly or annual plans while they consume content in minutes or at most hours. Thus, we approximate the duration of cost optimization to be of infinite horizon, over which the number of users (subscribers) is considered fixed.

A. User Demand Requirements

Over the infinite time horizon, each user n generates an independent sequence of requests $\{R_{n,t}\}_t$, where $R_{n,t} \in \{0, 1\}$ is an indicator of a request in slot t with $\pi_{n,t} := P(R_{n,t} = 1)$. That is, $\pi_{n,t}$ is the probability that user n generates a request at time t. We assume that to serve each request, the service provider consumes a uniform amount S of resources.[2] These demands are *delay-intolerant* in that when a request arrives to the service provider, it has to be fulfilled within the same time-slot of arrival. This is true for most content services of interest, e.g., on-demand video services, news, or social networking updates, especially under the large timescale network operation that is considered in our work.

B. Fluctuating Demand Pattern

Large timescale data networks are known to exhibit statistically fluctuating, *periodic* demand patterns, typically on a daily basis [4], [6], [11]. Accordingly, we assume that each day is divided into T time-slots whereby a user can generate one request per slot according to the statistics of $R_{n,t}$, and only require the following mild ergodicity characteristics:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} R_{n,l} = \overline{\pi}_n, \quad \text{w.p. } 1, \quad n = 1, \ldots N. \tag{1}$$

[1] *Delay-intolerant* in that the service must be received within the same slot that it is requested.

[2] We note that the results and insights obtained in this paper apply to the more general case of S being user-dependent, and some special cases of time dependency. In particular, $S_{n,t}$ is the amount of resources required to serve demand of user n at time t, with $S_{n,t}$ being known to the service provider, e.g., cyclic with some period T as discussed in the fluctuating demand case. Yet, we consider a constant S for simplicity of notation and ease of exposition.

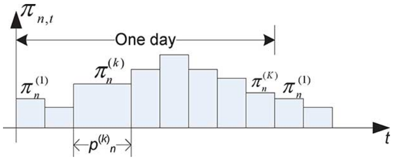


Fig. 1. Cyclostationary demand for user n. Average demand consistently assumes K different values every day depending on the user activity.

That is, $\overline{\pi}_n$ is the *time-average* demand probability for user n. Note that this assumption does not preclude the demand pattern from being time-varying, as will be introduced next.

The daily user activity is assumed to yield $K \geq 1$ levels of average demand. Each user n changes his demand probabilities through $\pi_n^{(1)}, \pi_n^{(2)}, \ldots, \pi_n^{(K)}$, over the course of the day. The demand probability of period k for user n spans a fraction $p_n^{(k)}$ of the T-slot day. That is, if the day starts by slot 0, then user n requests data w.p. $\pi_{n,t} = \pi_n^{(1)}$ on slots $t \in \{0, \ldots, p_n^{(1)}T - 1\}$, and so on, as illustrated in Fig. 1. Hence, we have $p_n^{(k)} \geq 0$, and $\sum_{k=1}^{K} p_n^{(k)} = 1$, $\forall n$. These demand characteristics are repeated consistently every day in accordance with the regular user activities as in Fig. 1.

In the sequel (especially in Section IV), we will use the compact notation $\mathbf{\Pi}^{(K)} := (\pi_n^{(k)}, p_n^{(k)})_{n=1,\ldots,N}^{k=1,\ldots,K}$ to characterize the fluctuating daily demand profile of the network.

C. Service Cost Structure

We assume that the cost of serving a total amount $x \geq 0$ of demand in a single slot is captured by a *strictly convex, increasing* function $C(x) : \mathbb{R}_+ \to \mathbb{R}$. Minimization of time-averaged costs for such functions calls for the smoothing of the load over time, as is desired by all service providers. In our numerical investigations, we will consider polynomial forms for this cost, while the results are obtained for the above general class. Thus, the obtained results in this work hold under convex cost structure. As we consider large timescale of operation whereby time-slot duration spans several minutes, fading dynamics of wireless channel are assumed averaged out, and hence are not incorporated in the cost function. Similar assumptions have been considered in other works on large timescale optimization such as [11]–[15].

D. Reactive Operation Paradigm

As a baseline scenario, we consider the predominant practice of *reactive* network operation, whereby the requests are served upon their arrival. Thus, under reactive service, the total load present at the service provider in slot t is given by: $L_t^{\mathcal{R}} := S \sum_{n=1}^{N} R_{n,t}$, since all N user requests initiated in slot t have to be served in the same slot. Thus, the corresponding time-average expected cost for the reactive service model is given by $c^{\mathcal{R}}(\mathbf{\Pi}^{(K)}) := \lim_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \mathbb{E}\left[C\left(L_l^{\mathcal{R}}\right)\right]$, where the distribution of $\{R_{n,t}\}_{n,t}$ is governed by the profile $\mathbf{\Pi}^{(K)}$ as described above. Clearly, such a reactive model represents a worst-case cost performance for the service provider side as it carries no proactive resource allocation strategies.

E. Proactive Operation Paradigm

We assume that the service provider is aware of the demand profile $\mathbf{\Pi}^{(K)}$ that captures the statistical characteristics of future demand, yet with uncertainties (see Fig. 2).

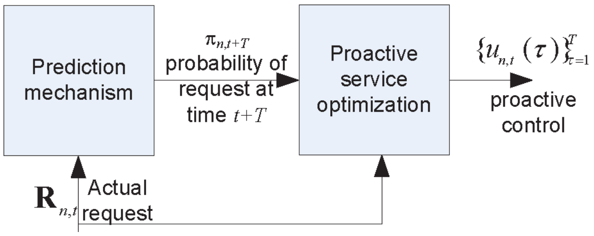

Prediction mechanism
$\pi_{n,t+T}$
probability of request at time $t+T$
Proactive service optimization
$\{u_{n,t}(\tau)\}_{\tau=1}^{T}$
proactive control
$\mathbf{R}_{n,t}$
Actual request

Fig. 2. Proactive service model.

Based on such uncertainties, proactive data services are carefully employed over a day-ahead (T-slot) time window so as to smooth out the network traffic over time. We denote by $u_{n,t}(\tau)$ the amount of service applied at time t to a potential request from user n that is expected to arrive after τ time-slots, i.e., at time $t+\tau$, where $1 \leq \tau \leq T$. The proactive service of a future request can not exceed the total demand of S units of service, i.e.,

$$\sum_{\tau=1}^{T} u_{n,t-\tau}(\tau) \leq S \qquad \forall n, t \tag{2}$$

and the proactive service can never be negative, i.e.,

$$u_{n,t}(\tau) \geq 0 \qquad \forall n, t, \tau. \tag{3}$$

Then, the expected system load in a time-slot t under proactive control $\mathbf{u}_t := [u_{n,t}(\tau)]_{n,\tau}$ is given by

$$L_t^{\mathcal{P}}(\mathbf{u}_t) := \sum_{n=1}^{N} \left(\left(S - \sum_{\tau=1}^{T} u_{n,t-\tau}(\tau) \right) R_{n,t} + \sum_{\tau=1}^{T} u_{n,t}(\tau) \right)$$

which consists of the on-time service component resulting from the nonproactively served part of the request, and the proactive service of future requests in the upcoming T-slot interval (compare to $L^{\mathcal{R}}$ under the reactive operation).

Here, we note the implicit assumption that the service provider is only uncertain about whether each user is going to generate a request or not, i.e., uncertainty about user activity. Yet, the service provider is assumed to anticipate the exact content the user will demand given the user generates a request at all. This assumption is motivated by the recent advances on machine learning, collaborative filtering, and big data analysis that enable several service providers (e.g., Netflix and YouTube) to successfully recommend content to subscribers [27]–[29].

In addition, our earlier works [25], [26] have particularly studied the impact of uncertainty about the exact content to be requested by end-users, and have established the notion of demand shaping through valuations and pricing incentives whereby service providers can significantly enhance such certainty and quality of proactive downloads. Thus, in this work we move on to the study of the impact of uncertainty about the user activity assuming service providers are capable of fully predicting the exact content to be requested, in case of a request.

F. Problem Statement and Notion of Asymptotic Optimality

The objective of proactive design is to develop the controller that minimizes the time average expected cost

$$c_T^{\mathcal{P}}(\mathbf{\Pi}^{(K)}) := \min_{\{\mathbf{u}_l\}_l} \quad \limsup_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \mathbb{E}\left[C\left(L_l^{\mathcal{P}}(\mathbf{u}_l, \mathbf{\Pi}^{(K)})\right)\right]$$
$$\text{s.t.} \quad \text{Constraints (2), (3)} \tag{4}$$

where the subscript T captures the proactive service window size,[3] and the superscript $\mathcal{P}$ indicates proactive operation.

The exact solution of (4) is intractably complex due to the infinite dimensionality of the problem. Instead, we aim to develop efficient proactive caching policies that can optimally utilize statistical predictions as the prediction window grows. Fortunately, our analysis will also show that the performance of these policies converge to the optimal exponentially fast, thereby possess close-to-optimal performance even for moderate values of the prediction window.

Before we present the design and analysis of such asymptotically optimal stationary policies, we formally define the notion of asymptotic optimality as follows.

Definition 1: A proactive caching policy $\mathbf{p}$ is *asymptotically optimal* under the demand profile $\mathbf{\Pi}^{(K)}$ if $\limsup_{T\to\infty} |c_T^{\mathbf{p}}(\mathbf{\Pi}^{(K)}) - c_T^{\mathcal{P}}(\mathbf{\Pi}^{(K)})| = 0$.

III. Proactive Service of Uniform Demand

We break down our analysis into two scenarios: that of *uniform demand* discussed in the current section, and that of *fluctuating demand* postponed to Section IV. Uniform demand means we have time-invariant prediction errors of future demand, whereas fluctuating demand means that uncertainties are time-varying according to a cyclostationary pattern as observed in datasets [4], [6]. This is done for two reasons. First and foremost, considering uniform demand allows us to isolate the impact of prediction uncertainties from that of fluctuations. Second, the uniform demand case allows us to present the main approach without the notational complexity that time-varying demands necessitate.

Under our $\mathcal{U}$*niform Demand* model, $\{R_{n,t}\}_t$ for user n is an independent and identically distributed (i.i.d.) sequence of random variables with $\mathbb{E}[R_{n,t}] = \overline{\pi}_n$, i.e., all requests of the same user are statistically indistinguishable over time. Note that such uniform demand is a special case of the fluctuating pattern $\mathbf{\Pi}^{(K)}$ introduced in Section II with $K = 1$, $p_n^{(1)} = 1$ and $\pi_n^{(1)} = \overline{\pi}_n$, for all n. To clarify the distinction, throughout this section, we simply use $\overline{\boldsymbol{\pi}}$ instead of $\mathbf{\Pi}^{(K)}$ to characterize the uniform demand, and return to the fluctuating case in Section IV.

We remark that uniform demand, as opposed to a fluctuating demand with the same time-average, promises less proactive gains, as all time-slots being equally uncertain creates the highest confusion about the best way of proactively performing services. Therefore, the proactive gains under uniform demand comes only from the uncertain knowledge of the demand captured by $\overline{\boldsymbol{\pi}} := (\overline{\pi}_n)_n$.

Next, we establish a global lower bound as a function of the prediction uncertainties $\overline{\boldsymbol{\pi}}$ on the minimum attainable cost by any proactive policy, and investigate its characteristics. In Section III-A, we will use the lower bound to develop an asymptotically optimal proactive caching policy.

A. Lower Bound on Minimum Cost for Uniform Demand

In this section, we first establish a lower bound on the optimal performance by any proactive caching policy under uniform demand. We draw interesting insights and remarks on the impact of uncertainty on optimal proactive caching by contrasting the resulting bound with that achievable only with infinite delay tolerance, and that achieved by reactive operation.

Theorem 1 (Lower Bound for Uniform Demand): Let $\mathcal{B}_t =: \{n \in \mathcal{N} : R_{n,t} = 1\}$ be the set of users that generate data requests at time t according to $\overline{\boldsymbol{\pi}} = (\overline{\pi}_n)_n$. Then, under uniform demand, and for any $T \geq 1$, the optimal proactive caching cost, $c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}})$ of (4), satisfies

$$c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \geq \underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}}) \tag{5}$$

$$\begin{aligned} \underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}}) := \min_{\{\tilde{\mu}_n(\mathcal{B})\}_{n,\mathcal{B}}} & \left\{ \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B}) \times \right. \\ & \left. C\left(\sum_{n\in\mathcal{B}} \left(S - \sum_{\mathcal{D}\subseteq\mathcal{N}} P_1(\mathcal{D})\tilde{\mu}_n(\mathcal{D}) \right) + \sum_{n=1}^{N} \tilde{\mu}_n(\mathcal{B}) \right) \right\} \\ \text{s.t.} \quad & 0 \leq \tilde{\mu}_n(\mathcal{B}) \leq S \qquad \forall n, \mathcal{B} \end{aligned} \tag{6}$$

where $P_1(\mathcal{B}) := \prod_{n\in\mathcal{B}} \overline{\pi}_n \prod_{m\notin\mathcal{B}} (1 - \overline{\pi}_m)$ is the probability of set $\mathcal{B}_t = \mathcal{B}$ under the uniform demand model.

Proof: Please refer to Appendix A. ■

In the objective of (6), the term $\sum_{\mathcal{D}\subseteq\mathcal{N}} P_1(\mathcal{D})\tilde{\mu}_n(\mathcal{D})$ captures the average proactive service assigned to a request from user n before it is actually realized, where $\mathcal{D} \subseteq \mathcal{N}$ is a possible set of requesting users, and the term $\tilde{\mu}_n(\mathcal{B})$ is the total proactive service assigned to all possible requests from user n when the current set of demanding users is $\mathcal{B}$. The theorem establishes that no proactive caching policy can achieve a lower cost than the nontrivial bound $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}})$ under the uniform demand model. We note that the optimization of $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}})$ is convex and yields a unique solution by the strict convexity of $C(\cdot)$. Such optimization is numerically tractable and can be easily computed, e.g., through dual or interior-point methods.

The bound $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}})$ is interesting in its own right, as it captures the impact of unavoidable prediction uncertainties $\overline{\boldsymbol{\pi}}$ on the lowest attainable cost by proactive design, even if it is known infinitely ahead of time. Deferring the goal of attaining this bound to Section III-B, we next develop some interesting insights on it.

Insights on the Lower Bound: 1) We first contrast our bound $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}})$ to the trivial lower bound under *infinitely* delay-tolerant services given by $c^*(\overline{\boldsymbol{\pi}}) := C\left(S \sum_{n=1}^{N} \overline{\pi}_n\right)$. It is known that this $c^*(\overline{\boldsymbol{\pi}})$ level of cost is achievable by stationary policies as the delay tolerance grows to infinity (see, e.g., [31] in the context of smart grids). However, it is loose for proactive services of *delay-intolerant* demands with *unavoidable prediction uncertainties* $\overline{\boldsymbol{\pi}}$. This is because proactive caching must experience the costs of prediction uncertainties that can be eliminated by delay-tolerant services at the expense of (potentially unboundedly high) delay. In fact, we next prove that proactive caching cannot achieve $c^*(\overline{\boldsymbol{\pi}})$ unless at the extreme conditions where the prediction is perfect.

Theorem 2 (Unavoidable Costs of Uniform Uncertainty): Under the uniform demand model with given $\overline{\boldsymbol{\pi}}$, $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}}) \geq c^*(\overline{\boldsymbol{\pi}})$, with equality if and only if $\overline{\pi}_n \in \{0, 1\}, \forall n \in \mathcal{N}$.

Proof: We first establish the result that full certainty about future demand is necessary to achieve $c^*(\overline{\boldsymbol{\pi}})$.

Lemma 1: Let $G_1(t) := \frac{1}{t}\sum_{l=0}^{t-1}\sum_{n=1}^{N}\sum_{\tau=1}^{T} u_{n,l}(\tau)$, and $G_2 := \frac{1}{t}\sum_{l=0}^{t-1}\sum_{n=1}^{N}\sum_{\tau=1}^{T} u_{n,l-\tau}(\tau)R_{n,l}$, then a proactive

[3] There is a slight abuse of notation since T represents both the number of slots per day and the proactive window size. That is, proactive service can be applied up to one day ahead.

caching policy $\{\mathbf{u}_t\}_t$ asymptotically achieves $c^*(\bar{\boldsymbol{\pi}})$ only if $\liminf_{t\to\infty} G_1(t) - G_2(t) = 0$ w.p.1.

Proof: The proof follows by Jensen's inequality and Fatou's Lemma. Please refer to Appendix D. ■

Back to the Proof of Theorem 2: Note that $G_1(t)$ represents the average proactive service applied, whereas $G_2(t)$ is the average amount of such service that is actually matched by user demand, hence made useful for the service provider. The difference $G_1(t) - G_2(t)$ captures the wasted proactive service due to future demand uncertainties.

Lemma 1 shows that $\liminf_{t\to\infty} G_1(t) - G_2(t) = 0$ w.p. 1 is necessary to have the equality hold for any prediction window T. Yet, for the uniform demand model and $t > T$, we have

$$\sum_{l=0}^{t-1}\sum_{n=1}^{N}\sum_{\tau=1}^{T} u_{n,l}(\tau) \geq \sum_{l=0}^{t-1}\sum_{n=1}^{N}\sum_{\tau=1}^{T} u_{n,l-\tau}(\tau) R_{n,l}, \quad \text{w.p. } 1$$

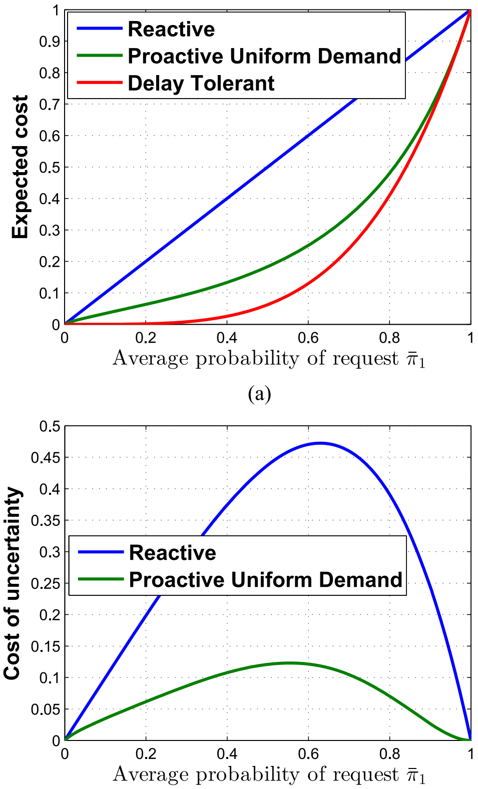
with equality if and only if $R_{n,l}$ is identically 1 or 0 on $l = 0, 1, \ldots$ for all n. This is realized when $\bar{\pi}_n$ is either 1 or 0, respectively. Note that, when $R_{n,l}$ is identically 0 on $l = 0, \ldots$, the optimal control $u_{n,l}(\tau)$ is trivially 0 on $l = 0, 1, \ldots$. ■

2) The previous insight shows that no proactive policy can attain the delay-tolerant cost of $c^*(\bar{\boldsymbol{\pi}})$ except when $\bar{\pi}_n \in \{0, 1\}$ for all n. Now, we turn to understanding the nature of $\underline{c}_{\mathcal{U}}(\bar{\boldsymbol{\pi}})$ within the extremes. The value of $\bar{\pi}_n$ in such model captures all information about demand uncertainty as well as average demand level. We present some key insights on the impact of $\bar{\boldsymbol{\pi}}$ on the lower bound, in particular how it affects the proactive gains.

Consider the proactive service of a single-user requesting S units of service in each slot with probability $\bar{\pi}_1$. The service of x units of service in a slot incurs a polynomial cost function of degree $d > 1$, that is $C(x) = x^d$.

Fig. 3(a) contrasts, for the case of $d = 4$, $S = 1$, and increasing values of $\bar{\pi}_1$, the average costs $c^{\mathcal{R}}(\bar{\pi}_1)$, $\underline{c}_{\mathcal{U}}^{\mathcal{P}}(\bar{\pi}_1)$, and $c^*(\bar{\pi}_1)$ achievable, respectively, by the reactive, proactive, and infinitely delay-tolerant schemes. It shows that proactive caching attains considerably lower cost as compared to the reactive one, and almost follows the same trend of the delay-tolerant case. In fact, the impact of uncertainty can be seen through the slope of the cost curve. Proactive caching cost increases slowly at small values of $\bar{\pi}_1$ as the system best utilizes the increasing certainty as well as low load levels. Yet, as $\bar{\pi}_1$ increases, costs tend to increase faster as the system becomes more congested with incoming requests, thereby restricting the chances of shifting the load. To further highlight the proactive-reactive comparison, in Fig. 3(b), we plot the *cost of uncertainty* of reactive and proactive operation by plotting $(c^{\mathcal{R}}(\bar{\boldsymbol{\pi}}) - c^*(\bar{\boldsymbol{\pi}}))$ and $(\underline{c}_{\mathcal{U}}(\bar{\boldsymbol{\pi}}) - c^*(\bar{\boldsymbol{\pi}}))$, which measures the relative cost incurred due to lack of exact information about future demand, as compared to that of delay-tolerant services. It reveals both reactive and proactive schemes must suffer the cost of uncertainty with the extremes of $\bar{\pi}_1 \in \{0, 1\}$. Yet, reactive scheduling suffers heavy cost of uncertainty since it does not utilize the statistical information about future demand, however uncertain it is. In contrast, proactive caching suffers significantly less as it exploits the statistical knowledge about future.

3) While uniform demand scenario assumes $\pi_{n,t} = \bar{\pi}_n, \forall t$, the lower bound $\underline{c}_{\mathcal{U}}(\bar{\boldsymbol{\pi}})$ still applies if $\{\pi_{n,t}\}_{n,t}$ are unknown (unobservable) random variables, yet satisfy the ergodicity condition (1). This can be seen through the proof of Theorem 1, where conditioning over set $\mathcal{B}_l$ is still applicable, and ergodicity condition (1) ensures that $P(\mathcal{B}_l = \mathcal{B}) = P_1(\mathcal{B})$. Thus, systems that are unaware of the per-slot demand probability $\pi_{n,t}$ (even with $\pi_{n,t}$ are nonidentical over time) cannot attain cost performance that is smaller than $\underline{c}_{\mathcal{U}}(\bar{\boldsymbol{\pi}})$.



(a)

(b)

Fig. 3. Comparison of reactive, proactive, and delay-tolerant costs under the uniform demand pattern. (a) Average incurred cost. (b) Cost of uncertainty.

We now move on to the design and analysis of a specific proactive caching policy that asymptotically achieves the lower bound $\underline{c}_{\mathcal{U}}(\bar{\boldsymbol{\pi}})$ with growing prediction window size.

B. Stationary Proactive Caching Policy Design and Analysis

In this section, we introduce a simple proactive caching policy for the uniform demand model, prove its asymptotic optimality as defined in Definition 1, establish its convergence rate, and analyze its performance gains compared to its reactive counterpart for a specific setting.

Definition 2 (Proactive Caching Policy $\mathbf{P}_{\mathcal{U}}$): Given the observed requests $\mathcal{B}_t = \{n \in \mathcal{N} : R_{n,t} = 1\}$ in slot t, our proactive caching policy $\mathbf{p}_{\mathcal{U}}$ sets its proactive control parameter as: $u_{n,t}(\tau) = \frac{\mu_n(\mathcal{B}_t)}{T}, \forall n, t, \tau$, where $\{\mu_n(\mathcal{B})\}_{n,\mathcal{B}}$ is the optimal solution of the minimization in (6).

Policy $\mathbf{p}_{\mathcal{U}}$, thus, is a stationary policy that observes $\mathcal{B}_t$, the set of users who request content at time t, and accordingly assigns proactive control value $u_{n,t}(\tau) = \mu_n(\mathcal{B}_t)/T$ for all potential requests to may be requested in the upcoming T slots. Intuitively, this policy determines its proactive download amounts in response to $\mathcal{B}_t$ by utilizing the solution of the lower bound (6), and then mimics *processor-sharing* type of service discipline

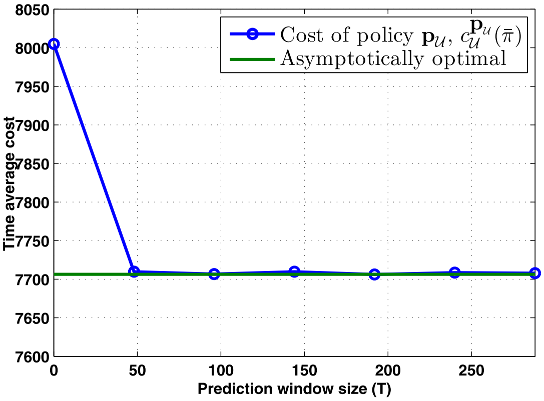


Fig. 4. Impact of proactive window size on achievable cost.

by equally spreading the given amount over the horizon of the prediction window T. Next, we show that this policy is asymptotically optimal as T increases.

Theorem 3 (Asymptotic Optimality): Under the uniform demand pattern described by $\overline{\boldsymbol{\pi}}$, our proactive caching policy $\mathbf{p}_{\mathcal{U}}$ is asymptotically optimal (cf. Definition 1), therefore it achieves $\underline{c}_{\mathcal{U}}(\overline{\boldsymbol{\pi}})$ as $T \to \infty$.

Proof: Please refer to Appendix B. ■

While asymptotic optimality is theoretically encouraging, it is of practical interest to find out whether the policy possess desirable performance guarantees in nonasymptotic regimes. This motivates us, next, to quantify the speed at which $\mathbf{p}_{\mathcal{U}}$ performance approaches its asymptotic limit as the prediction window size T grows.

Theorem 4 (Exponential Bounds on Convergence Speed): Let $\delta > 0$, then there exists a function $g_1 > 1$ such that $g_1(\delta) \to 1$ as $\delta \to 0$, and

$$c_T^{\mathbf{p}_{\mathcal{U}}}(\overline{\boldsymbol{\pi}}) - c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \leq \delta M + 2^N M g_1(\delta)^{-T} \tag{7}$$

where M is a positive constant.

Proof: Please refer to Appendix C. ■

This theorem establishes the desirable nonasymptotic property of our policy $\mathbf{p}_{\mathcal{U}}$ that its cost reaches an arbitrarily small neighborhood of the optimal achievable level exponentially fast with T. Such a property can be attributed to the processor-sharing nature of the control variables (e.g., $\mu_n(\mathcal{B})/T$ being assigned to all prospective requests), which enables traffic averaging over time, thanks to strong law of large numbers. Hence, system randomness decays exponentially with the prediction window size.

In Fig. 4, we plot the achieved time-average cost under $\mathbf{p}_{\mathcal{U}}$ against the prediction window size T to show its rapid convergence to established lower bound. In the simulation, $N = 15$, $\overline{\pi}_1 = 0.5933$, $C(L) = L^4$.

Before concluding this section, we also share some insights on the question of when the cost reduction of proactive operation relative to its reactive counterpart is greatest. We consider the scenario (as in Fig. 3) of a single-user requesting S units of service in each slot with probability $\overline{\pi}_1$. The service of x units of service in a slot incurs a polynomial cost function of degree $d > 1$, that is $C(x) = x^d$. In this scenario, policy $\mathbf{p}_{\mathcal{U}}$ reduces to $u_{1,t}(\tau) = \mu(0)/T$ if $R_{1,t} = 0$, and $\mu(1)/T$ if $R_{1,t} = 1$, where

$$(\mu(0), \mu(1)) := \arg \min_{(\tilde{\mu}(0), \tilde{\mu}(1)) \succcurlyeq \mathbf{0}}$$
$$\overline{\pi}_1(S - (1 - \overline{\pi}_1)\tilde{\mu}(0) - \overline{\pi}_1\tilde{\mu}(1) + \tilde{\mu}(1))^d + (1 - \overline{\pi}_1)(\tilde{\mu}(0))^d.$$

Through simple differentiation, we get that $\mu(1) = 0$ and $\mu(0) = \sqrt[d-1]{\overline{\pi}_1} S / (1 + (1 - \overline{\pi}_1) \sqrt[d-1]{\overline{\pi}_1})$.

Proposition 1: For the above single-user model with costs $C(L) = L^d$, the relative cost reduction of our policy $\mathbf{p}_{\mathcal{U}}$ with respect to reactive performance function calculated as $\gamma(\overline{\pi}_1) := \frac{c^{\mathcal{R}}(\overline{\pi}_1) - \lim_{T \to \infty} c_T^{\mathbf{p}_{\mathcal{U}}}(\overline{\pi}_1)}{c^{\mathcal{R}}(\overline{\pi}_1)}$ is maximized at the unique value of $\overline{\pi}_1^* = 1/d$, with the value $\gamma(\overline{\pi}_1^*) = \frac{1}{S^d}\left(\frac{d^{\frac{1}{d-1}+1} S}{d^{\frac{1}{d-1}+1}+d-1}\right)^d + \frac{(1-d)}{S^d}\left(\frac{d\, S}{d^{\frac{1}{d-1}+1}+d-1}\right)^d + 1$, with $\gamma(\overline{\pi}_1^*) \to 1$ as $d \to \infty$.

The proof of proposition follows easily from simple calculus while noting that $\mu(0)$ is a function of $\overline{\pi}_1$. Proposition 1 clearly shows that $\overline{\pi}_1 = 1/d$ is the best operating point for proactive gains over reactive ones under the uniform demand pattern and polynomial cost function. It is clear that similar optimal operating points will arise under different cost functions, sharing the common characteristic of optimally balancing between utilizing the certainty about future demand, and creating the opportunity for proactive services in idle slots. On the other hand, the relative cost reduction achieved asymptotically by an infinitely delay-tolerant operation is given by $(1 - \overline{\pi}_1^{d-1})$, which is monotonically decreasing in $\overline{\pi}_1$, as there is no waste of service, hence increasing $\overline{\pi}_1$ can only limit potential caching opportunities and reduce the system gain.

Remark 1 (Complexity of $\mathbf{p}_{\mathcal{U}}$): The controls of policy $\mathbf{p}_{\mathcal{U}}$ require the solution to (6). While such optimization has significant number of variables, $N2^N$, it needs only to be solved once, offline, based on the long-term system statistics $\overline{\boldsymbol{\pi}}$. Then, in the online operation, the service provider has to only observe $\mathcal{B}_t$ and use the mapping $\{\mu_n(\mathcal{B}_t)\}_n$ for proactive service. The size of such mapping is 2^N, yet using *binary* search, the complexity of determining proactive controls in any time-slot under policy $\mathbf{p}_{\mathcal{U}}$ is $O(N)$.

IV. Proactive Service of Fluctuating Demand

In this section, we return to the general model of *$\mathcal{F}$luctuating Demand* (i.e., $K > 1$) characterized by $\mathbf{\Pi}^{(K)} := (\pi_n^{(k)}, p_n^{(k)})_{n=1,\dots,N}^{k=1,\dots,K}$ as described in Section II, and study the performance of proactive caching strategies that utilize such fluctuations to proactively shift traffic forward in time and attain minimum service costs. The development follows the same structure as in Section III but with heavier notation due to the time-varying statistics of the demands.

A. Lower Bound on Minimum Cost for Fluctuating Demand

Similar to Section III, we begin by introducing a lower bound on the minimum time-average cost achievable by any proactive caching policy, and contrast its performance for varying $\mathbf{\Pi}^{(K)}$ to the infinitely delay-tolerant and reactive costs.

Recall that the probability of demand from user n at time t is given by $\pi_{n,t}$. Accordingly, under fluctuating demand, we can write

$$\pi_{n,t} = \begin{cases} \pi_n^{(1)}, & t \pmod T < p^{(1)}T \\ \pi_n^{(k)}, & \sum_{m=1}^{k-1} p_n^{(m)} T \leq t \pmod T < \sum_{m=1}^{k} p^{(m)} T. \end{cases}$$

We collect the demand probabilities of all users at time t in a set $\mathcal{J}_t$ as follows. $\mathcal{J}_t := \{\pi_{n,t}\}_{n=1}^N$. Finally, we quantify the fraction of daily time-slots through which set $\mathcal{J}$ of demand probabilities is realized by $Q_2(\mathcal{J})$, which is given as[4]

$$Q_2(\mathcal{J}) = \begin{cases} 0, & \sum_{m=1}^{k_1} p_{\nu_1}^{(m)} < \sum_{m=1}^{k_2-1} p_{\nu_2}^{(m)}, \\ & \text{for any } p_{\nu_1}^{(k_1)}, p_{\nu_2}^{(k_2)} \in \mathcal{J}, k_2 > 1 \\ \min_n p_n^{(1)}, & \mathcal{J} = \{p_n^{(1)}\}_n \\ \min_n p_n^{(K)}, & \mathcal{J} = \{p_n^{(K)}\}_n \\ 1 - \check{J} - \hat{J}, & \text{otherwise} \end{cases}$$

where $\check{J} := \min_{(n,k)}\{\sum_{m=1}^{k} p_n^{(m)} : p^{(k+1)} \in \mathcal{J}, k > 1\}$, $\hat{J} := \min_{(n,k)}\{\sum_{m=k+1}^{K} p_n^{(m)} : p^{(k-1)} \in \mathcal{J}, k < K\}$.

Now, we are ready to present the general lower bound in the following theorem.

Theorem 5 (Lower Bound for Fluctuating Demand): Let $\mathcal{K} := \{\pi_1^{(k)}\}_{k=1}^{K} \times \cdots \times \{\pi_N^{(k)}\}_{k=1}^{K}$ be the set of all N-tuple demand probabilities for the N users, and $\mathbf{\Pi}^{(K)} = (\pi_n^{(k)})_{n,k}$ characterizes complete fluctuating demand profile. Then, for any $T \geq 1$, the optimal proactive caching cost $c_T^{\mathcal{P}}(\mathbf{\Pi}^{(K)})$ satisfies

$$\begin{aligned} & c_T^{\mathcal{P}}(\mathbf{\Pi}^{(K)}) \\ & \geq \underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)}) \qquad (8) \\ & \underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)}) \\ & := \min_{\{\tilde{\mu}_n(\mathcal{B},\mathcal{I},\mathcal{J})\}_{\mathcal{B},\mathcal{I},\mathcal{J}}} \Bigg\{ \sum_{\mathcal{I}\in\mathcal{K}} \sum_{\mathcal{B}\subseteq\mathcal{N}} Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) \times \\ & \quad C\Bigg(\sum_{n\in\mathcal{B}}\Bigg(S - \sum_{\mathcal{J}\in\mathcal{K}} \sum_{\mathcal{D}\subseteq\mathcal{N}} Q_2(\mathcal{J}) P_2(\mathcal{D}|\mathcal{J}) \tilde{\mu}_n(\mathcal{D},\mathcal{J},\mathcal{I})\Bigg) \\ & \quad + \sum_{n=1}^{N} \sum_{\mathcal{J}\in\mathcal{K}} Q_2(\mathcal{J}) \tilde{\mu}_n(\mathcal{B},\mathcal{I},\mathcal{J})\Bigg)\Bigg\} \\ & \quad \text{s.t.} \quad 0 \leq \tilde{\mu}_n(\mathcal{B},\mathcal{I},\mathcal{J}) \leq S, \quad \forall n, \mathcal{B} \subseteq \mathcal{N}, \quad \mathcal{I},\mathcal{J} \in \mathcal{K} \\ & P_2(\mathcal{B}|\mathcal{I}) \\ & := P(\mathcal{B}_t = \mathcal{B}|\mathcal{J}_t = \mathcal{I}) \\ & = \prod_{\{(n,k):n\in\mathcal{B},\pi_n^{(k)}\in\mathcal{I}\}} \pi_n^{(k)} \prod_{\{(m,l):m\notin\mathcal{B},\pi_m^{(l)}\in\mathcal{I}\}} (1 - \pi_m^{(l)}). \qquad (9) \end{aligned}$$

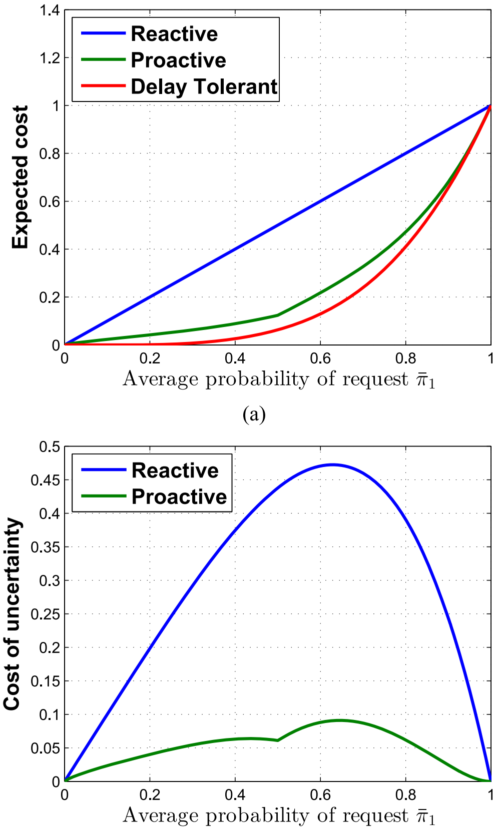
Proof: Follows the same steps as that of Theorem 1, except conditioning goes over $(\mathcal{B}_t, \mathcal{J}_t)$ instead of $\mathcal{B}_t$ only. Hence, it is omitted for brevity. ■

Note that the lower bound in (9) is essentially more sophisticated than (5) due to the additional information available at the service provider that differentiates between the demand over daily time-slots. Yet, the optimization is still tractable as the problem is convex.

Next, similar to Theorem 2, we establish that $\underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)})$, achieves the delay-tolerant limit $c^*(\mathbf{\Pi}^{(K)}) = C(S\sum_{n=1}^{N}\overline{\pi}_n)$ if and only if demand prediction approaches full certainty.

Theorem 6 (Unavoidable Costs of Fluctuating Uncertainty): Under the fluctuating demand pattern with given $\mathbf{\Pi}^{(K)}$, $\underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)}) \geq c^*(\mathbf{\Pi}^{(K)})$, with equality if and only if $\pi_n^{(k)} \in \{0,1\}$ for all n, k.

Proof: Please refer to Appendix E. ■



(b)

Fig. 5. Comparison of reactive, proactive, and delay-tolerant costs under sinusoidally fluctuating demands. (a) Average incurred cost. (b) Cost of uncertainty.

In Fig. 5, we numerically compare the lower bound $\underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)})$ to the minimum costs achievable by reactive and infinitely delay-tolerant schemes. In this setup, we set $N = 1$, $K = 12$, $C(L) = L^4$, and set $\pi_1^{(k)} = \overline{\pi}_1 + \min\{\overline{\pi}_1, 1-\overline{\pi}_1\}\sin(\frac{2\pi(k-1)}{K})$, which idealistically captures the peak–off-peak characteristics of the daily demand pattern through a sinusoidal function. As the average load $\overline{\pi}_1$ varies from 0 to 1, we observe from Fig. 5(a) that proactive costs stay very closely to the idealized delay-tolerant limit, while the reactive costs perform very poorly. Also, Fig. 5(b) illustrates the cost of uncertainty with the same metric as in Fig. 3(b). We can clearly see from the figure that fluctuating demands offer further gains over uniform demands since fluctuations enable shifting of the load to less congested durations with smaller service costs.

B. Cyclostationary Proactive Caching Policy Design and Analysis

We start by defining a cyclostationary proactive service policy for fluctuating demand patterns.

Definition 3 (Proactive Caching Policy $\mathbf{P}_{\mathcal{F}}$): Given the observed requests $\mathcal{B}_t = \{n \in \mathcal{N} : R_{n,t} = 1\}$ in slot t, our proactive caching policy $\mathbf{p}_{\mathcal{F}}$ selects its proactive service amounts as: $u_{n,t}(\tau) = \frac{\mu_n(\mathcal{B}_t,\mathcal{J}_t,\mathcal{J}_{t+\tau})}{T}$, for each $\tau \in \{1,\ldots,T\}$ where $\{\mu_n(\mathcal{B},\mathcal{I},\mathcal{J})\}_{\mathcal{B},\mathcal{I},\mathcal{J}}$ is the unique solution to (9).

Here, we note that policy $\mathbf{p}_{\mathcal{F}}$ does not only assign proactive services based on the current realization of user requests $\mathcal{B}_t$, but also it incorporates the statistical information about the current demand $\mathcal{J}_t$, as well as future demand $\mathcal{J}_{t+\tau}$, $\tau = 1,\ldots,T$, in

[4] Note that $\sum_{\mathcal{J}\in\mathcal{K}} Q_2(\mathcal{J}) = 1$.

its decisions so as to maximally utilize the available resources at minimum cost. Clearly, the policy is cyclostationary with period T since $\mathcal{J}_t = \mathcal{J}_{t+mT}$ for any positive integer m. In the following theorem, we establish the asymptotic optimality of $\mathbf{p}_{\mathcal{F}}$.

Theorem 7 (Asymptotic Optimality): Under the fluctuating demand pattern described by $\mathbf{\Pi}^{(K)}$, our proactive caching policy $\mathbf{p}_{\mathcal{F}}$ is asymptotically optimal (cf. Definition 1), therefore it achieves $\underline{c}_{\mathcal{F}}(\mathbf{\Pi}^{(K)})$ as $T \to \infty$.

Proof: Under a cyclostationary policy $\mathbf{p}$, the resulting average cost can be expressed as $c_T^{\mathbf{p}}(\mathbf{\Pi}^{(K)}) = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[C(L_t^{\mathcal{P}}(\mathbf{u}_t))\right]$. By expanding the expectation through conditioning on $(\mathcal{B}_t, \mathcal{J}_t)$, and following similar steps to the proof of Theorem 3, the result follows. ■

Hence, we can see that $\sum_{\mathcal{J}\in\mathcal{K}}\sum_{\mathcal{D}\subseteq\mathcal{N}} Q_2(\mathcal{J}) \times P_2(\mathcal{D}|\mathcal{J})\mu_n(\mathcal{D},\mathcal{J},\mathcal{I})$ is the average proactive service received by user n at the time instants with demand probabilities of all users form set $\mathcal{I}$. Furthermore, the term $\sum_{\mathcal{J}\in\mathcal{K}} Q_2(\mathcal{J})\mu_n(\mathcal{B},\mathcal{I},\mathcal{J})$ captures the proactive services assigned to user n when $\mathcal{B}$ is current set of requesting users, $\mathcal{I}$, is the current set of demand levels, and $\mathcal{J}$ is the potential set of demand levels at which a request from user n is expected to be realized.

As policy $\mathbf{p}_{\mathcal{F}}$ employs a processor sharing discipline for its proactive services, similar to $\mathbf{p}_{\mathcal{U}}$ it also enjoys an exponential converging speed with T to the ultimate lower bound (8), as established next.

Theorem 8 (Exponential Bounds on Convergence Speed): Let $\delta > 0$, then there exists a function $g_2 > 1$ such that $g_2 \to 1$ as $\delta \to 0$ and

$$c_T^{\mathbf{p}_{\mathcal{F}}}(\mathbf{\Pi}^{(K)}) - c_T^{\mathcal{P}}(\mathbf{\Pi}^{(K)}) \le \delta M + 2^N M g_2(\delta)^{-T}$$

for some positive constant M.

In Fig. 6, we explicitly plot the achieved time-average cost against the prediction window size T under fluctuating demand pattern with characteristics specified as follows. There are $N = 15$ users in the system who request services from a random fashion. The day to be divided into $K = 24$ h with the average probability of demand for each user n varies over the course of the day according to (0.73, 0.73, 0.73, 0.78, 0.73, 0.73, 0.78, 0.86, 0.90, 0.90, 0.78, 0.67, 0.49, 0.31, 0.31, 0.2, 0.2, 0.2, 0.36, 0.43, 0.54, 0.54, 0.67, 0.67). Here, the first element $\pi_n^{(1)} = 0.73$ corresponds to the time period (00 am, 1 am], and the last element $\pi^{(24)} = 0.67$ corresponds to the period (11 pm, 00 am]. Thus, $\bar{\pi}_n = \frac{1}{K}\sum_{k=1}^{K}\pi_n^{(k)} = 0.5933$, $\forall n$. We adopt T as the number of daily time-slots. Thus, in the simulation, $T = 288$ corresponds to a slot size of 5 min, which is reasonable for a user to generate one data request.

To further highlight the impact of the proposed policy $\mathbf{p}_{\mathcal{F}}$ on the system's load under different values of T, we show in Fig. 7 the daily average load levels achieved by both reactive and proactive caching, including two instances of T in addition to the asymptotically optimal limit. Clearly, as T grows, load levels become smoother over time, and at $T = 288$ (corresponding to a time-slot size of 5 min) the proactive cost is indistinguishable from the asymptotic optimal. We also observe that proactive caching considerably smooths out the load over time, while uncertainty still yields some minor fluctuations that cannot be avoided according to Theorem 5. Here, we note that different service providers can divide the day into different number of time-slots T. For instance, Netflix may operate at $T = 24$ slots with slot duration being an hour as it serves movies with long duration. CNN, on the other hand, can divide the day into larger number of time-slots, e.g., $T = 288$ slots, since its videos are shorter.

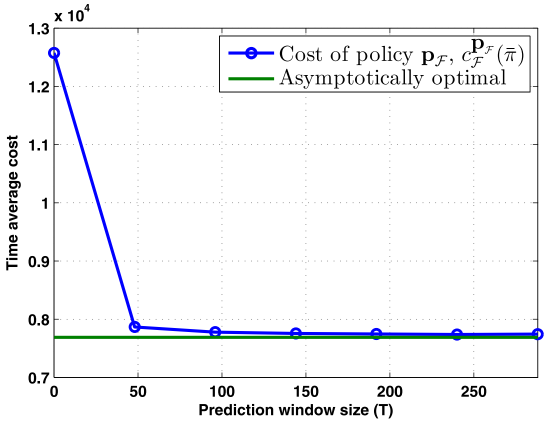


Fig. 6. Impact of proactive window size on achievable cost of policy $\mathbf{p}_{\mathcal{F}}$ for fluctuating demand.

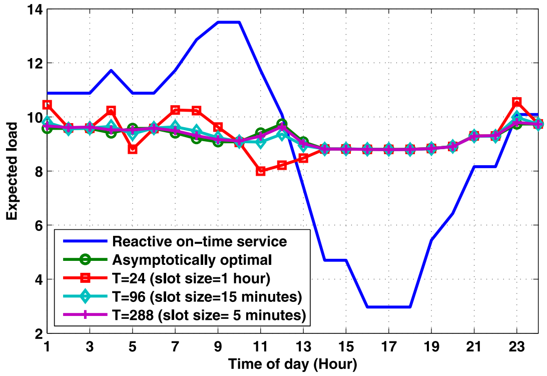


Fig. 7. Average load levels under reactive and proactive services for fluctuating demands.

Remark 2 (Complexity of $\mathbf{P}_{\mathcal{F}}$): Similar to policy $\mathbf{p}_{\mathcal{U}}$, optimization (9) needs to only be solved once, offline, based on long-term characteristics $\mathbf{\Pi}^{(K)}$. Then, in the online operation, mapping $\mu_n(\mathcal{B}_t, \mathcal{J}_t, \mathcal{J}_{t+\tau})$ is harnessed to assign proactive control $u_{n,t}(\tau)$. The complexity of determining proactive controls under $\mathbf{p}_{\mathcal{F}}$ in any time-slot is $O(N + \log H)$,[5] where $H = \sum_{\mathcal{J}\in\mathcal{K}} \mathbf{1}_{\{Q_2(\mathcal{J})\}}$, and $\mathbf{1}_{\{x\}} = 1$, if $x > 0$, and $\mathbf{1}_{\{x\}} = 0$ otherwise.

Remark 3 (Memory Consumption): To enable proactive caching at end-users, there is an amount $\sum_{\tau=1}^{T} u_{n,t}(\tau)$ of data cached on device every time-slot t. On the other hand, an amount $\sum_{\tau=1}^{T} u_{n,t-\tau}(\tau)$ of previously applied proactive service becomes irrelevant at the end of time-slot t and thus can be overwritten. As a result, dynamic memory allocation for proactive service of user n is given by $\sum_{i=1}^{T}\sum_{\tau=i}^{T} u_{n,t-i+1}(\tau)$. Proactive downloads are also chosen to decrease linearly as the size of the set of requesting users $|\mathcal{B}|$ grows.

[5] It is also possible to develop efficient low-complexity solutions to (6) and (9) with high performance guarantees. Nevertheless, this is not the main focus of this work and hence can be addressed separately.

In the light of Remark 3, expected memory usage per slot under policy $\mathbf{p}_{\mathcal{U}}$ for uniform demand is given by $\frac{T+1}{2}\sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B})\mu_n(\mathcal{B})$. Furthermore, expected memory usage under policy $\mathbf{p}_{\mathcal{F}}$ for fluctuating demand is upper-bounded by $\frac{T+1}{2}A_n$, where $A_n := \max_{\mathcal{I}\in\mathcal{K}} \sum_{\mathcal{J}\in\mathcal{K}} \sum_{\mathcal{D}\subseteq\mathcal{N}} Q_2(\mathcal{J})P_2(\mathcal{D}|\mathcal{J})\mu_n(\mathcal{D},\mathcal{J},\mathcal{I})$.

It is clear that memory requirements grow with prediction window size. Nevertheless, the exponential convergence of operational cost to the established lower bounds (see Theorems 4 and 8) suggests moderate values of T will yield a best balance between memory allocation and operational cost. In addition, for fair comparison to delay-tolerant networks, we note that despite such memory allocation requirements, proactive caching does not suffer any service delays and hence enhances quality of experience.

In the numerical simulations above, we have noted that choice of $T = 288$ slots corresponds to slot size of 5 min. Now, for this case, the upper bound on memory allocation under fluctuating demand is $50S$. An average 5-min YouTube video has a size of $\sim$ 10 MB. Thus, memory allocation for proactive caching in such case will be less than 0.5 GB. As recent versions of smartphones support large storage (e.g., 128 GB), it is clear that memory requirements for proactive service are well satisfied.

Remark 4 (Memory Constraints): From the above discussion, expected memory consumption at every time-slot for user n can be expressed as an affine expression of $\{\mu_n(\mathcal{B})\}_{\mathcal{B}}$ for uniform demand, and $\{\mu_n\mathcal{B},\mathcal{I},\mathcal{J}\}_{\mathcal{B},\mathcal{I},\mathcal{J}}$ for fluctuating demand. In case that users pose memory constraints on proactive service, lower bound optimizations (6), (9) can be modified to include such new affine constraints, where convexity is preserved. However, the lower bound will be dependent on proactive window size T, that is, for each value of T, we have a lower bound on minimum achievable cost. Nevertheless, our proposed policies $\mathbf{p}_{\mathcal{U}}, \mathbf{p}_{\mathcal{F}}$ are still valid and can be applied to smooth out traffic over time.

V. More Numerical Results

In this section, we provide additional numerical results to reap further insights on the performance of the proposed proactive caching policies.

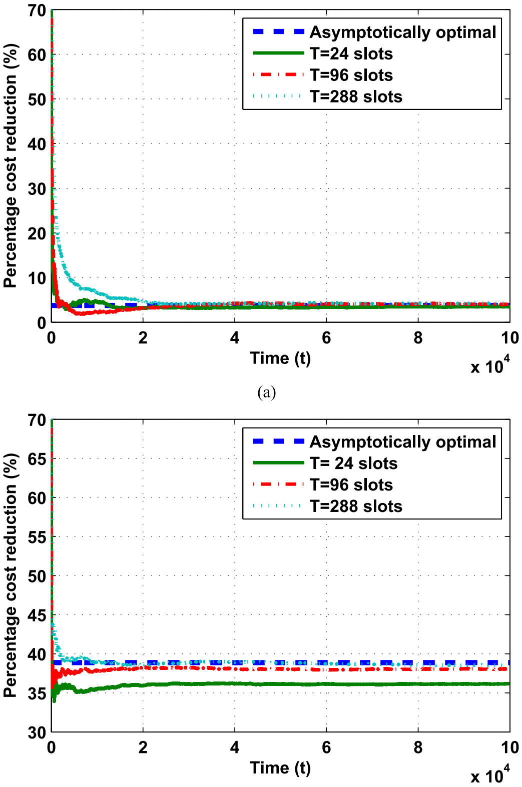
A. Time-Domain Performance

Following the exact simulation setup used in Figs. 3 and 4, we plot in Fig. 8 the evolution of relative cost reduction gain with time for both models of uniform and fluctuating demand patterns through the use of policies $\mathbf{p}_{\mathcal{U}}, \mathbf{p}_{\mathcal{F}}$, respectively.

For both models of uniform and fluctuating demand, increasing T enhances the attainable cost reduction gain until it approaches the respective upper bound obtained as T grows. We also note that the worst-case prediction under uniform demand, resulting from identically distributed requests over time, essentially limits the system gains due to significant uncertainty. However, for fluctuating demand, the system is able to attain remarkably higher gains by exploiting the time variability of demand levels.

B. Impact of Number of Users

We study the impact of the number of users on the system's performance for both models of demand patterns (uniform and fluctuating) in Fig. 9. In particular, we plot the *asymptotically optimal* cost reduction gain against the number of users N, We also consider the asymptotic performance of a DTN with full certainty about generated requests. We can see that performance of uniform demand [Fig. 9(a)] considerably falls with N, even for the DTN with full certainty. The reason is that the increased randomness of the system because of more users, together with the statistically indistinguishable requests, limits the opportunities of shifting the demand over time. On the other hand, fluctuating demand attains considerably higher gains (although non-increasing with N) by leveraging major caching opportunities offered in off-peak hours. Here, we note that *the comparison above does not show the significant user dissatisfaction associated with the DTN due to large delays, which is completely resolved via proactive caching.*



(a)

(b)

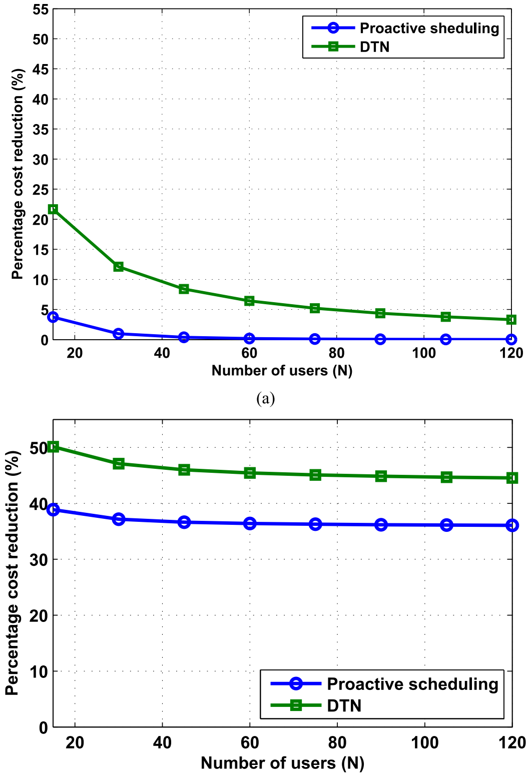
Fig. 8. Convergence of relative cost reduction with time. (a) Policy $\mathbf{p}_{\mathcal{U}}$ for uniform demand. (b) Policy $\mathbf{p}_{\mathcal{F}}$ for fluctuating demand.

C. Certainty-Cost Reduction Tradeoff

Finally, in Fig. 10, we highlight the tradeoff between uncertainty and relative cost reduction gain for the simple example of $N = 1$ user addressed in Section III. Since the cost function is polynomial with degree $d = 4$, it is clear that $\overline{\pi}_1 = 1/4$ is the optimal value of $\overline{\pi}_1$ that strikes the best balance between certainty, and enough opportunities for load shift over time. The DTN counterpart, on the other hand, attains a monotonically decreasing gain with $\overline{\pi}_1$ since it does not suffer any uncertainty issues; only increasing $\overline{\pi}_1$ decreases the opportunity for exploiting empty slots for load balancing.

VI. Conclusion

In this work, we have considered the notion of *proactive* caching of delay-intolerant data services in the presence of uncertain predictions of future user demands. We consider service



(a)

(b)

Fig. 9. Impact of the number of users N on achievable cost. (a) Policy $\mathbf{p}_{\mathcal{U}}$ for uniform demand. (b) Policy $\mathbf{p}_{\mathcal{F}}$ for fluctuating demand.

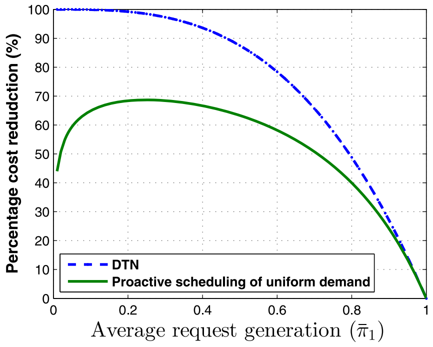


Fig. 10. Relative cost reduction gain versus $\overline{\pi}_1$ for uniform demand.

providers that utilize the statistically predictable nature of future user demands to selectively serve data requests before their actual time of realization. We revealed that by harnessing time instants with low demand characteristics, service providers can leverage significant cost reduction through proactive service of future requests. Despite the ultimate challenge of uncertainty about future requests, considerably lower service costs have been proven achievable via proactive caching in large timescale systems where delay tolerance is essentially limited. We have studied interesting instances of predictable demand, established fundamental lower bounds on the achievable costs through proactive caching, and developed asymptotically optimal policies that attain these bounds rapidly as the proactive caching window size increases. Furthermore, we have contrasted the asymptotically optimal performance with that of infinitely deferrable ideal scenario, and drawn interesting insights and remarks on the unavoidable costs of prediction uncertainties.

Appendix A
Proof of Theorem 1

Let $\{\mathbf{u}_t^*\}_t$ be an optimal proactive caching policy under the uniform demand model, where $\mathbf{u}_t^* = [u_{n,t}^*(\tau)]_{n,\tau}$. Thus, we can write

$$c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) = \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\Big[C\big(L_l^{\mathcal{P}}(\mathbf{u}_l)\big)\Big].$$

By conditioning on all possible sets of requesting users at time $l \geq 0$, we can rewrite this as

$$c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \\ = \limsup_{t\to\infty} \frac{1}{t}\sum_{l=1}^{t-1}\sum_{\mathcal{B}\subseteq\mathcal{N}} P(\mathcal{B}_l=\mathcal{B})\mathbb{E}\Big[C\big(L_l^{\mathcal{P}}(\mathbf{u}_l)\big)|\mathcal{B}_l=\mathcal{B}\Big].$$

By Jensen's inequality, since C is assumed strictly convex, we have

$$\begin{aligned} c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \geq \limsup_{t\to\infty} \frac{1}{t}\sum_{l=1}^{t-1}\sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B}) \\ \times C\Bigg(\sum_{n\in\mathcal{B}}\Bigg(S-\sum_{\tau=1}^{T}\mathbb{E}[u_{n,l-\tau}^*(\tau)]\Bigg) \\ +\sum_{n=1}^{N}\Bigg(\sum_{\tau=1}^{T}\mathbb{E}[u_{n,l}^*(\tau)|\mathcal{B}]\Bigg)\Bigg). \end{aligned}$$

Note that $\{\mathcal{B}_l\}_l$ is an i.i.d. sequence under the uniform demand pattern, thus we could use $P_1(\mathcal{B})$. Moreover, $\mathcal{B}_l$ is independent of $\sum_{\tau=1}^{T} u_{n,l-\tau}^*(\tau)$, $l \geq 0$.

Since $\frac{1}{t}\sum_{l=0}^{t-1} 1 = 1$, we can apply Jensen's inequality again to have

$$\begin{aligned} & c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \\ & \geq \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B})C\Bigg(\sum_{n\in\mathcal{B}}\Bigg(S-\liminf_{t\to\infty}\frac{1}{t}\sum_{l=1}^{t-1}\sum_{\tau=1}^{T}\mathbb{E}[u_{n,l-\tau}^*(\tau)]\Bigg) \\ & \quad +\sum_{n=1}^{N}\Bigg(\limsup_{t\to\infty}\frac{1}{t}\sum_{l=1}^{t-1}\sum_{\tau=1}^{T}\mathbb{E}[u_{n,l}^*(\tau)|\mathcal{B}]\Bigg)\Bigg). \end{aligned}$$

As C is monotonically increasing, we can replace $\limsup$ on the right-hand side (RHS) of the last expression by $\liminf$. Furthermore, by defining $\tilde{\mu}_n(\mathcal{B}) := \liminf_{t\to\infty}\frac{1}{t}\sum_{l=0}^{t-1}\sum_{\tau=1}^{T}\mathbb{E}[u_{n,l}^*(\tau)|\mathcal{B}]$, we obtain

$$\begin{aligned} c_T^{\mathcal{P}}(\overline{\boldsymbol{\pi}}) \geq \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B})C\Bigg(\sum_{n\in\mathcal{B}}\Bigg(S-\sum_{\mathcal{D}\subseteq\mathcal{N}}P_1(\mathcal{D})\tilde{\mu}_n(\mathcal{D})\Bigg) \\ +\sum_{n=1}^{N}\Big(\tilde{\mu}_n(\mathcal{B})\Big)\Bigg). \end{aligned}$$

Note that Constraints (2) and (3) imply that $0 \leq \tilde{\mu}_n(\mathcal{B}) \leq S$, $\forall n, \mathcal{B}$. Now, by minimizing the right-hand side of the last expression over all feasible choices of $\{\tilde{\mu}_n(\mathcal{B})\}_{\mathcal{B}}$, the theorem is proved.

Appendix B
Proof of Theorem 3

It suffices to prove that $\limsup_{T\to\infty} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) = \liminf_{T\to\infty} c_T^{\mathcal{P}}(\boldsymbol{\pi})$. We start by $\limsup_{T\to\infty} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi})$. Since $\mathbf{p}_\mathcal{U}$ is a stationary policy that depends only on the current demand realization, we can write

$$\begin{aligned} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) &= \mathbb{E}\left[C\left(\sum_{n=1}^{N}\left(S-\sum_{\tau=1}^{T-1}u_{n,t-\tau}(\tau)\right)R_{n,t}+\sum_{\tau=1}^{T}u_{n,t}(\tau)\right)\right] \\ &= \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B})\mathbb{E}\left[C\left(\sum_{n\in\mathcal{B}}\left(S-\sum_{\tau=1}^{T}u_{n,t-\tau}(\tau)\right)\right.\right. \\ &\qquad\qquad \left.\left. +\sum_{n=1}^{N}\mu_n(\mathcal{B})\right)|\mathcal{B}_t=\mathcal{B}\right]. \end{aligned}$$

Now, we consider the sum $\sum_{\tau=1}^{T} u_{n,t-\tau}(\tau)$, which is independent of $\mathcal{B}_t$. Define a counter $\mathbf{Z}_T(\mathcal{D})$ that measures the number of occurrences of a requesting set of users $\mathcal{D}\subseteq\mathcal{N}$ in slots $t-T,\ldots,t-1$. Then, $\sum_{\tau=1}^{T}u_{n,t-\tau}(\tau)=\sum_{\mathcal{D}\subseteq\mathcal{N}}\frac{\mu_n(\mathcal{D})\mathbf{Z}_T(\mathcal{D})}{T}$. Thus, by the strong law of large numbers, as $T\to\infty$

$$\limsup_{T\to\infty}\sum_{\mathcal{D}\subseteq\mathcal{N}}\frac{\mu_n(\mathcal{D})\mathbf{Z}_T(\mathcal{D})}{T}=\sum_{\mathcal{D}\subseteq\mathcal{N}}\mu_n(\mathcal{D})P_1(\mathcal{D}),\quad \text{w.p. 1.}$$

By noting that the system load $L_t^{\mathcal{P}}(\mathbf{u}_t)\le 2NS$, bounded convergence theorem implies

$$\begin{aligned} \limsup_{T\to\infty} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) &= \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B}) \\ \times C&\left(\sum_{n\in\mathcal{B}}\left(S-\sum_{\mathcal{D}\subseteq\mathcal{N}}P_1(\mathcal{D})\mu_n(\mathcal{D})\right)+\sum_{n=1}^{N}\mu_n(B)\right). \end{aligned} \tag{10}$$

Second, by noting that the RHS of (10) is identical to $\underline{c}_\mathcal{U}(\boldsymbol{\pi})$, then $\limsup_{T\to\infty} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) \le \liminf_{T\to\infty} c_T^{\mathcal{P}}(\boldsymbol{\pi})$. Yet, by the definition of $c_T^{\mathcal{P}}(\boldsymbol{\pi})$, it follows that $\limsup_{T\to\infty} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) = \liminf_{T\to\infty} c_T^{\mathcal{P}}(\boldsymbol{\pi})$, which completes the proof.

Appendix C
Proof of Theorem 4

Let $\mathbf{Y}_\mathcal{D} := \frac{1}{T}\sum_{\tau=1}^{T}\mathbf{1}_{\{\mathcal{D}_{t-\tau}=\mathcal{D}\}}$, where $\mathbf{1}_A$ is the indicator function of event A. We have

$$\begin{aligned} P\left(|\mathbf{Y}_\mathcal{D}-P_1(\mathcal{D})|>\delta\right) &= P\left(\mathbf{Y}_\mathcal{D}>P_1(\mathcal{D})+\delta\right) \\ &\quad +P\left(\mathbf{Y}_\mathcal{D}<P_1(\mathcal{D})-\delta\right). \end{aligned}$$

From Chernoff bound, we can write

$$P\left(\mathbf{Y}_\mathcal{D}>P_1(\mathcal{D})+\delta\right)\le\inf_{r>0}e^{T(\Lambda(r)-r(P_1(\mathcal{D})+\delta))}$$

where $\Lambda(r)$ is the log moment generating function of $\mathbf{1}_{\{\mathcal{D}_t=\mathcal{D}\}}$, which is a Bernoulli random variable with parameter $P_1(\mathcal{D})$. Hence, the tightest Chernoff bound is attained at $r^*=\log(1+\frac{\delta}{P_1(\mathcal{D})(1-P_1(\mathcal{D})-\delta)})$, which yields

$$\begin{aligned} P(\mathbf{Y}_\mathcal{D}>P_1(\mathcal{D})+\delta) &\le h_\mathcal{D}(\delta)^{-T},\quad \delta>0 \\ h_\mathcal{D}(\delta) &= \frac{\left(1+\frac{\delta}{P_1(\mathcal{D})(1-P_1(\mathcal{D})-\delta)}\right)^{(P_1(\mathcal{D})+\delta)}}{1+\frac{\delta}{1-P_1(\mathcal{D})-\delta}}. \end{aligned} \tag{11}$$

Note that $h_\mathcal{D}(\delta)>1$, $\delta>0$. Similarly, we can show that

$$P(\mathbf{Y}_\mathcal{D}<P_1(\mathcal{D})-\delta)\le h_\mathcal{D}(-\delta)^{-T},\quad \delta\in(0,P_1(\mathcal{D})). \tag{12}$$

Note also that $h_\mathcal{D}(-\delta)>1$ on $\delta\in(0,P_1(\mathcal{D}))$.

Now, we define $\mathcal{A}_\delta:=\{\mathbf{Y}_\mathcal{D}:|\mathbf{Y}_\mathcal{D}-P_1(\mathcal{D})|\le\delta\}$, thus we can expand and bound $c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi})$ as

$$\begin{aligned} c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi}) = \sum_{\mathcal{B}\subseteq\mathcal{N}} P_1(\mathcal{B})\Big\{&P(\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta,\forall\mathcal{D}) \\ &\times\mathbb{E}\Big[C\Big(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}\mathbf{Y}_\mathcal{D}\mu_n(\mathcal{D}) \\ &\qquad+\sum_{n}\mu_n(\mathcal{B})\Big)\Big|\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta,\forall\mathcal{D}\Big] \\ &+P(\mathbf{Y}_{\mathcal{D}'}\notin\mathcal{A}_\delta \text{ for some } \mathcal{D}')) \\ &\times\mathbb{E}\Big[C\Big(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}\mathbf{Y}_\mathcal{D}\mu_n(\mathcal{D}) \\ &\qquad+\sum_{n}\mu_n(\mathcal{B})\Big)\Big|\mathbf{Y}_{\mathcal{D}'}\notin\mathcal{A}_\delta,\text{ for some }\mathcal{D}'\Big]\Big\} \\ \le\sum_{\mathcal{B}\subseteq\mathcal{N}}P_1(\mathcal{B})&\left(C\left(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}(P_1(\mathcal{D})-\delta)\mu_n(\mathcal{D})\right.\right. \\ &\qquad\left.+\sum_{n=1}^{N}\mu_n(\mathcal{B})\right)P(\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta,\forall\mathcal{D}) \\ &\left.+C(2NS)P(\mathbf{Y}_{\mathcal{D}'}\notin\mathcal{A}_\delta\text{ for some }\mathcal{D}')\right). \end{aligned}$$

The inequality follows since $C(\cdot)$ is monotonically increasing, $\mathbf{Y}_\mathcal{D}=P_1(\mathcal{D})-\delta$ is the value of $\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta$ that maximizes the conditioned cost on $\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta$, and $2NS$ is the largest load the service provider can sustain under the constraints on $\{\mu_n(\mathcal{B})\}_{n,\mathcal{B}}$.

The difference $c_T^{\mathbf{p}_\mathcal{U}}(\boldsymbol{\pi})-c_T^{\mathcal{P}}(\boldsymbol{\pi})$ can be upper-bounded by

$$\begin{aligned} &\sum_{\mathcal{B}\subseteq\mathcal{N}}P_1(\mathcal{B})\left(\left(C\left(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}(P_1(\mathcal{D})-\delta)\mu_n(\mathcal{D})\right.\right.\right. \\ &\qquad\left.+\sum_{n=1}^{N}\mu_n(\mathcal{B})\right)-C\left(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}P_1(\mathcal{D})\mu_n(\mathcal{D})\right. \\ &\qquad\left.\left.+\sum_{n=1}^{N}\mu_n(\mathcal{B})\right)\right)P(\mathbf{Y}_\mathcal{D}\in\mathcal{A}_\delta,\forall\mathcal{D}) \\ &\qquad+\left(C(2NS)-C\left(\sum_{n\in\mathcal{B}}S-\sum_{\mathcal{D}\subseteq\mathcal{N}}P_1(\mathcal{D})\mu_n(\mathcal{D})\right.\right. \\ &\qquad\left.\left.\left.+\sum_{n=1}^{N}\mu_n(\mathcal{B})\right)\right)P(\mathbf{Y}_{\mathcal{D}'}\notin\mathcal{A}_\delta\text{ for some }\mathcal{D}')\right) \\ &\overset{(a)}{\le}\delta SC'(2NS)+SC'(2NS)P(\mathbf{Y}_{\mathcal{D}'}\notin\mathcal{A}_\delta\text{ for some }\mathcal{D}') \\ &\overset{(b)}{\le}\delta M+M2^N\max_{\mathcal{D}'\subseteq\mathcal{N}}P(|\mathbf{Y}_{\mathcal{D}'}-P_1(\mathcal{D}')|>\delta) \\ &\overset{(c)}{\le}\delta M+M2^N\max_{\mathcal{D}'\subseteq\mathcal{N}}h_{\mathcal{D}'}(-\delta)^{-T}+h_{\mathcal{D}'}(\delta)^{-T} \end{aligned}$$

where $C'(\cdot)$ is the first derivative of $C(\cdot)$, and $M=SC'(2NS)$. Inequality (a) follows by mean value theorem and monotonicity[6] of $C'(\cdot)$ since $C(X) - C(Y) \leq C'(X)(X - Y)$. Also, $P(\mathbf{Y}_{\mathcal{D}} \in \mathcal{A}_{\delta}, \forall \mathcal{D}) \leq 1$. Inequality (b) follows from upper-bounding $P(\mathbf{Y}_{\mathcal{D}'} \notin \mathcal{A}_{\delta}$, for some $\mathcal{D}')$ by $2^N \max_{\mathcal{D}' \subseteq \mathcal{N}} P(|\mathbf{Y}_{\mathcal{D}'} - P_1(D)| > \delta)$, which by (11) and (12) leads to inequality (c). Now, by setting $g_1(\delta) = \max_{\mathcal{D}' \subseteq \mathcal{N}} \ 2 \min\{h_{\mathcal{D}'}(\delta),\ h_{\mathcal{D}'}(-\delta)\}$, the proof is completed.

Appendix D
Proof of Lemma 1

We have by Jensen's inequality

$$c_T^{\mathcal{P}} \geq \min_{\{\mathbf{u}_l\}_l} \limsup_{t \to \infty} \mathbb{E}\left[C\left(\frac{1}{t} \sum_{l=0}^{t-1} L_l^{\mathcal{P}}(\mathbf{u}_l) \right) \right].$$

Note that C is a strictly convex by hypothesis, and the expectation operator preserves convexity. We can write

$$\sum_{l=0}^{t-1} L_l^{\mathcal{P}}(\mathbf{u}_l) = S \sum_{l=0}^{t-1} \sum_{n=1}^{N} R_{n,t} + \sum_{l=0}^{t-1} \sum_{n=1}^{N} \sum_{\tau=1}^{T} u_{n,l}(\tau) - \sum_{l=0}^{t-1} \sum_{n=1}^{N} \sum_{\tau=1}^{T} u_{n,l-\tau}(\tau) R_{n,l}. \quad (13)$$

Furthermore, since C is monotonically increasing, we have

$$\begin{aligned} c_T^{\mathcal{P}} &\geq \limsup_{t \to \infty} \mathbb{E}\left[C\left(\frac{1}{t} \sum_{l=0}^{t-1} S \sum_{n=1}^{N} R_{n,l} + G_1(t) - G_2(t) \right) \right] \\ &\overset{(a)}{\geq} \mathbb{E}\left[C\left(S \sum_{n=1}^{N} \liminf_{t \to \infty} \frac{1}{t} \sum_{l=0}^{t-1} R_{n,l} + G_1(t) - G_2(t) \right) \right] \\ &\overset{(b)}{=} c^*. \end{aligned}$$

In (a), we used Fatou's lemma to replace lim sup outside the expectation with lim inf inside it. In (b), we used the fact that $G_1(t) \geq G_2(t)$ for any $t \geq 0$. Hence, if equality holds, then $\liminf G_1(t) - G_2(t) = 0$ w.p. 1.

Appendix E
Proof of Theorem 6

($\Rightarrow$) WLOG, we assume that $\mathcal{N}$ is the set of users n with $\pi_n^{(k)} \in \{0, 1\}, \forall k$, and $\overline{\pi}_n \notin \{0, 1\}$. Clearly, in case of $\overline{\pi}_n \in \{0, 1\}$, it is optimal to have zero proactive service for such traffic. Now, we have

$$P_2(\mathcal{B}|\mathcal{I}) = \begin{cases} 1, & \text{if } \forall (n, \pi_n^{(k)}) : n \in \mathcal{B}, \pi_n^{(k)} \in \mathcal{I}, \pi_n^{(k)} = 1, \\ & \quad \forall (n, \pi_n^{(k)}) : n \notin \mathcal{B}, \pi_n^{(k)} \in \mathcal{I}, \pi_n^{(k)} = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $\mathcal{J}_t$ carries all information about $\mathcal{B}_t$, and therefore we can omit the dependence on $\mathcal{B}_t$ from the rest of the proof. Hence, we can write the general lower bound for **M2** as

$$\sum_{\mathcal{I} \in \mathcal{K}} Q_2(\mathcal{I}) C\left(\sum_{n : \pi^{(k)} n \in \mathcal{I}, \pi_n^{(k)} = 1} \left(S - \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{J}, \mathcal{I}) \right) + \sum_{n=1}^{N} \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{I}, \mathcal{J}) \right). \quad (14)$$

Choosing $\mu_n(\mathcal{I}, \mathcal{J}) = S$ if $\pi_n^{(k)} \in \mathcal{I}$, $\pi_n^{(k)} = 0$, and $\pi_n^{(m)} \in \mathcal{J}$, with $\pi_n^{(m)} = 1$. Furthermore, set $\mu_n(\mathcal{I}, \mathcal{J}) = 0$ otherwise. Then, expression (14) reduces to

$$C\left(\sum_{n=1}^{N} S \overline{\pi}_n \right) = c^*.$$

Note that $\sum_{\mathcal{I} \in \mathcal{K}} Q_2(\mathcal{I}) = 1$, $\sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_m(\mathcal{I}, \mathcal{J}) = \overline{\pi}_n S$ if $\pi_n^{(k)} \in \mathcal{I}$, and $\overline{\pi}_n^{(k)} = 0$ for some k, $\sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_m(\mathcal{I}, \mathcal{J}) = 0$ otherwise.

($\Leftarrow$) Suppose $\underline{c}_{\mathcal{F}}(\boldsymbol{\pi}_K) = c^*$. Then, by the convexity of C, we have

$$\sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} \left(\sum_{n \in \mathcal{B}} \left(S - \sum_{\mathcal{J} \in \mathcal{K}} \sum_{\mathcal{D} \subseteq \mathcal{N}} Q_2(\mathcal{J}) P_2(\mathcal{D}|\mathcal{J}) \mu_n(\mathcal{D}, \mathcal{J}, \mathcal{I}) \right) + \sum_{n=1}^{N} \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{B}, \mathcal{I}, \mathcal{J}) \right) Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) = \sum_{n=1}^{N} S \overline{\pi}_n.$$

Now, suppose towards contradiction that $\pi_{n_0}^{(k_0)} \in (0, 1)$ for some $n_0 \in \mathcal{N}$, $k_0 \in \{1, \ldots, K\}$. Consequently, for any $\mathcal{J}$ such that $\pi_{n_0}^{(k_0)} \in \mathcal{J}$, $P_2(\mathcal{D}|\mathcal{J}) \in (0, 1), \forall \mathcal{D} \subseteq \mathcal{N}$.

Since

$$\sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} \sum_{n \in \mathcal{B}} S Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) = \sum_{n=1}^{N} S \overline{\pi}_n$$

then we must have

$$\begin{aligned} \sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) \sum_{n=1}^{N} \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{B}, \mathcal{I}, \mathcal{J}) &= \sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) \\ &\quad \cdot \sum_{n \in \mathcal{B}} \sum_{\mathcal{J} \in \mathcal{K}} \sum_{\mathcal{D} \subseteq \mathcal{N}} Q_2(\mathcal{J}) P_2(\mathcal{D}|\mathcal{J}) \mu_n(\mathcal{D}, \mathcal{J}, \mathcal{I}). \end{aligned}$$

By rearranging terms, the last equality can be written as

$$\begin{aligned} \sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) \sum_{n=1}^{N} \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{B}, \mathcal{I}, \mathcal{J}) &= \sum_{\mathcal{I} \in \mathcal{K}} \sum_{\mathcal{B} \subseteq \mathcal{N}} Q_2(\mathcal{I}) P_2(\mathcal{B}|\mathcal{I}) \sum_{n=1}^{N} \sum_{\mathcal{J} \in \mathcal{K}} Q_2(\mathcal{J}) \mu_n(\mathcal{B}, \mathcal{I}, \mathcal{J}) \\ &\quad \cdot \sum_{\mathcal{D} \subseteq \mathcal{N} : n \in \mathcal{D}} P_2(\mathcal{D}|\mathcal{J}). \end{aligned}$$

Yet, by hypothesis we have $\sum_{\mathcal{J} \in \mathcal{K}} \sum_{\mathcal{D} \subseteq \mathcal{N} : n_0 \in \mathcal{D}} \ P_2(\mathcal{D}|\mathcal{J}) Q_2(\mathcal{J}) < 1$, which essentially contradicts the fulfillment of the above inequality. Note that $\sum_{\mathcal{D} : n_0 \notin \mathcal{D}} \ P_2(\mathcal{D}|\mathcal{J}) Q_2(\mathcal{J}) > 0$ since $\pi_{n_0}^{(k_0)} \in (0, 1)$.

References

[1] Cisco, "Cisco Visual Networking Index: Forecast and methodology 2012–2017,".

[2] FCC, "Spectrum Policy Task Force Report," FCC 02-155, 2002.

[3] FCC, "Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies, Notice of proposed rule making and order FCC 03-322, 2003.

[4] "About RRDtool," 2014 [Online]. Available: http://oss.oetiker.ch/rrdtool/

[5] "RRDtool gallery," 2015 [Online]. Available: http://oss.oetiker.ch/rrdtool/gallery/index.en.html

[6] The cost function $C(\cdot)$ is strictly convex and increasing, thus has a positive and monotonically increasing derivative $C'(\cdot)$.

[6] M. Feknous *et al.*, "Internet traffic analysis: A case study from two major European operators," in *Proc. IEEE ISCC*, Jun. 23–26, 2014, pp. 1–7.

[7] J. Mitola, III, "Cognitive RADIO: An integrated agent architecture for software defined radio," Doctor of Technology dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, 2000.

[8] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[9] D. Niyato and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of Nash equilibrium, and collusion," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, p. 192,202, Jan. 2008.

[10] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of network services," *IEEE/ACM Trans. Netw.*, vol. 8, no. 2, p. 171,184, Apr. 2000.

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

[12] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," in *Proc. IEEE INFOCOM WKSHPS*, Apr. 14–19, 2013, pp. 357–362.

[13] L. Xiaofeng, H. Pan, and P. Lio, "Offloading mobile data from cellular networks through peer-to-peer WiFi communication: A subscribe-and-send architecture," *China Commun.*, vol. 10, no. 6, pp. 35–46, Jun. 2013.

[14] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *Proc. IEEE INFOCOM WKSHPS*, Apr. 2013, pp. 351–356.

[15] A. Bar, D. Mimran, L. Chekina, Y. Elovici, and B. Shapira, "Nesto—Network selection and traffic offloading system for Android mobile devices," in *Proc. 9th IWCMC*, Jul. 1–5, 2013, pp. 337–342.

[16] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lect. Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.

[17] S. Shakkottai and A. Eryilmaz, "Optimization and control of communication networks," in *The Control Handbook*, W. S. Levine, Ed., 2nd ed. College Park, MD, USA: Univ. of Maryland Press, CRC Press, 2010.

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

[19] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. ACM SIGKDD KDD*, 2011, pp. 1100–1108.

[20] B. S. Jensen, J. E. Larsen, K. Jensen, J. Larsen, and L. K. Hansen, "Estimating human predictability from mobile sensor data," in *Proc. IEEE MLSP*, Sep. 2010, pp. 196–201.

[21] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4833–4854, Aug. 2013.

[22] L. Huang, S. Zhang, M. Chen, and X. Liu, "When backpressure meets predictive scheduling," 2013 [Online]. Available: http://arxiv.org/abs/1309.1110

[23] "Inmobly," [Online]. Available: http://www.inmobly.com

[24] "PAUL," [Online]. Available: http://www.paultheapp.com/

[25] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Trans. Netw.*, 2014, to be published.

[26] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Joint smart pricing and proactive content caching for mobile services," *IEEE/ACM Trans. Netw.*, 2015, to be published.

[27] J. Davidson *et al.*, "The YouTube video recommendation system," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 293–296.

[28] X. Amatriain and J. Basilico, "Netflix recommendations: Beyond the 5 stars," *Netflix Tech Blog*, Apr. 2012.

[29] S. Bhattacharjee, M. Bragin, and D. Zhdanov, "Accurate recommendations of online movie ratings: Large data sets with low dimensions and span of multiple years," in *Proc. Winter Conf. Business Intell.*, Salt Lake City, UT, USA, 2010.

[30] I.-H. Hou and R. Singh, "Capacity and scheduling of access points for multiple live video streams," 2013 [Online]. Available: http://arxiv.org/abs/1306.2360

[31] I. Koutsopoulos and L. Tassiulas, "Control and optimization meet the smart power grid—Scheduling of power demands for optimal energy management," 2010 [Online]. Available: http://arxiv.org/abs/1008.3614

John Tadrous (S'10–M'15) received the B.Sc. degree in electrical engineering from Cairo University, Cairo, Egypt, in 2008, the M.Sc. degree in wireless communications from Nile University, Giza, Egypt, in 2010, and the Ph.D. degree in electrical engineering from The Ohio State University, Columbus, OH, USA, in 2014.

He was a Research Assistant with the Wireless Intelligent Networks Center (WINC), 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 with the Information Processing Systems Lab, The Ohio State University, where he worked on proactive resource allocation and scheduling, smart data pricing, and information theory. Since 2014, he has joined the Center for Multimedia Communication (CMC), Rice University, Houston, TX, USA, as a Post-Doctoral Research Associate, where he works on modeling and analysis of interactive data traffic, full-duplex communications, and beamforming design for massive MIMO systems.

Atilla Eryilmaz (S'00–M'06) received the B.S. degree in electrical and electronics engineering from Boğaziçi University, Istanbul, Turkey, in 1999, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2001 and 2005, respectively.

Between 2005 and 2007, he worked as a Postdoctoral Associate with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently an Associate Professor of electrical and computer engineering with The Ohio State University, Columbus, OH, USA, where he has been a faculty member since 2007. His 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.

Dr. Eryilmaz 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 the IEEE/ACM TRANSACTIONS ON NETWORKING since 2015. He was a coauthor 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.