2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, October 2-5, 2018

On Optimal Proactive Caching with Improving Predictions over Time

John Tadrous*, Atilla Eryilmaz†
* Department of Electrical and Computer Engineering, Gonzaga University, WA, E-mail: tadrous@gonzaga.edu
† Department of Electrical and Computer Engineering, Ohio State University, OH, E-mail: eryilmaz.2@osu.edu

***Abstract*—This paper considers optimal proactive caching when future demand predictions improve over time as expected to happen in most prediction systems. In particular, our model captures the correlated demand pattern that is exhibited by end users as their current activity reveals progressively more information about their future demand. It is observed in previous work that, in a network where service costs grow superlinearly with the traffic load and static predictions, proactive caching can be harnessed to flatten the load over time and minimize the cost. Nevertheless, with time varying prediction quality, a tradeoff between load flattening and accurate proactive service emerges.**

In this work, we formulate and investigate the optimal proactive caching design under time-varying predictions. Our objective is to minimize the time average expected service cost given a finite *proactive* service window. We establish a lower bound on the minimal achievable cost by any proactive caching policy, then we develop a low complexity caching policy that strikes a balance between load flattening and accurate caching. We prove that our proposed policy is *asymptotically* optimal as the proactive service window grows. In addition, we characterize other non-asymptotic cases where the proposed policy remains optimal. We validate our analytical results with numerical simulation and highlight relevant insights.

I. Introduction

Proactive content caching has been proposed to offer enhanced quality of experience (QoE) while maximally utilizing the data network's resources [1]. The idea hinges on exploiting the recent advances in human behavioral modeling, machine learning, and collaborative filtering in predicting future user demand and serving it ahead of time. Proactive content caching has promised significant network gains both for end-users and content service providers (SPs). For an end user, receiving content before demand reduces service delays [1], [2] and the end user receives data content at *lower* prices compared to reactive service techniques [3], [4], [5]. On the other hand, SPs benefit from proactive content caching in regulating their traffic load over time, limiting the large disparity between peak and off-peak hour demand levels, and hence minimizing operational costs [6], [7].

Several works have investigated design strategies for proactive content caching from different perspectives. In [2], [8], proactive caching has been considered from a queuing theory perspective where *perfectly* predictable data requests are enqueued and served based on the remaining time to arrival. In [6], the impact of uncertainty about future demand has been addressed, order-optimal *static* proactive caching policies have been developed, and the notion of offering incentives to the end user for higher predictability has been introduced. In [4], [5], joint pricing incentives and proactive caching strategies have been considered but for a static setup. *Dynamic* proactive caching policies with demand uncertainty have then been developed in [7] under the assumption of time-invariant predictions. The concept of proactive caching has been extended beyond pushing future demand to the end user's device to caching at small base stations in [9], [10].

In all previous work, there has been no rigorous attempt to consider the impact of temporal correlation of the end user's demand on the proactive caching design even though it is a natural feature of human behavior. For instance, after the user has watched a two-hour movie on Netflix, they may not watch another one before a while. On the other hand, if the user runs the first song from a YouTube playlist, it is more likely that the next song will play out after the current one has finished. Such correlation in demand patterns offers significant information about the future which, if judiciously harnessed, boosts the quality of proactive caching decisions. Nevertheless, demand correlation raises significant complexity in the design of optimal online caching strategies due to the associated time dependence of the caching decisions on the future predictions. In some existing literature like [9], demand correlation has been mentioned, however caching decisions have been developed for a short-term optimization without considering a model that factors in the correlated user behavior itself.

In this work, we take the step of generalizing the user demand profile from being independent over time to a correlated one. In particular, we consider proactive content caching from the perspective of a SP that aims at minimizing its time average service costs while serving content to a user with correlated demand activity. The user correlation model is assumed to be a Markov process through which the user alternates between requesting content and idling. To the best of our knowledge, this is the first time such correlated user-SP interactions have been captured in online proactive caching policy design. Our contributions along with the paper outline can be summarized as follows:

• In Section II, we formulate the time average cost minimization problem in which the end-user generates a time correlated demand pattern and the SP optimizes its proactive caching control over a finite window of time slots, called

This work is supported by NSF grants: CCSS-EARS-1444026, CNS-NeTS-1514260, CNS-NeTS-1717045, CMMI-SMOR-1562065, CNS-ICN-WEN-1719371, and CNS-SpecEES-1824337; the DTRA grant HDTRA1-15-1-0003; HDTRA1-18-1-0050; and the QNRF Grant NPRP 7-923-2-344.

978-1-5386-6596-1/18/$31.00 ©2018 IEEE

proactive service window.

• In Section III, we establish a fundamental lower bound on the achievable cost by any proactive caching policy under our correlated demand model.

• In Section IV, we develop an asymptotically-optimal policy that is proven to achieve the lower bound as the proactive service window size grows.

• In Section V, we show that our proposed policy is *optimal* if the proactive service window is just one slot in length, or if the user idles for at least one slot after receiving the requested content, a typical characteristic of interactive apps like Yelp, web-browsing, online gaming, etc.

• We validate our analytical results with numerical simulations in Section VI and conclude our work in Section VII.

Finally, we note that, while our approach of lower bounding the achievable cost and developing an asymptotically optimal policy that achieves the bound has been first used in [7], we stress that this paper is fundamentally different. The correlation aspect of demand creates a tradeoff between smoothing out the traffic through early-on proactive caching and waiting as close as possible to the actual demand instant for maximal utilization of information about such demand. The tradeoff essentially calls for new techniques in both the lower bound construction and the proposed policy design.

II. System Model and Problem Formulation

We consider a time-slotted system that comprises a service provider (SP) that supplies data content to an end user. In this model, we are *not* confined to a particular timescale for the time slot duration as has been the case in [7]. For example, in the large timescale (order of minutes or hours) the service can be a video file, the time slot is the duration to watch such a video, and the SP can be YouTube, CNN, Netflix, etc. In the medium timescale (order of seconds) the service can be a server's response to a user-generated query in an interactive session [11], e.g., the SP can be Yelp sending a burst of packets containing relevant information to the user. In the small timescale (order of milliseconds or less) the service can be a single packet sent from a general SP to a device.

The user demand activity at time slot t is captured by a binary random variable $R_t \in \{0,1\}$, $t = 0, 1, \cdots$, where $R_t = 1$ only when the user sends a request and $R_t = 0$ when user is idle in slot t. The activity process $\{R_t\}_t$ forms a Markov chain (c.f. Fig. 1) with transition probabilities $P(R_t = j | R_{t-1} = i) = p_{ij}$, $i, j \in \{0,1\}$. Let π denote the steady state probability of generating a request in a time slot, then $\pi = \frac{p_{01}}{p_{10}+p_{01}}$.

The service of a data request consumes an amount S of the SP's available resources[1] which incurs a cost $C(L)$ on consuming an amount L of resources in any time slot, with C being a monotonically increasing and strictly convex function. The sharper C is the more the need for proactive caching.

The SP can proactively serve potential user requests before they are actually sent in order to regulate its resource consumption load (*load* for short) over time and minimize the operational cost. Nevertheless, since the data content may be updated frequently, the user discards service that is more than T slots old, i.e., if the user sends a request at slot t, any service that has been *proactively* cached in earlier than $t - T$ is discarded by the user.

Let $u_t(\tau)$ denote the amount of resources used by the SP in slot t to proactively cache the service of a probable request at slot $t + \tau$, $\tau = 1, \cdots, T$. As such, the total service load experienced by the SP in a given slot t is given by

$$L_t := (S - \sum_{\tau=1}^{T} u_{t-\tau}(\tau)) R_t + \sum_{\tau=1}^{T} u_t(\tau), \tag{1}$$

whereby the first term measures the "reactive" load to complete the service of the demand arriving in slot t and the second term captures the proactive service of the predictable future demand.

In networks where the SP *reactively* responds to generated user demand and does not apply any proactive service, the time average expected cost is given by $c_{\text{rea}} := \pi C(S)$.

We formulate the problem of minimizing the SP's time average expected cost when it utilizes the predictability of the future user activity along with the content freshness window T in proactive service as follows. Given the per-slot-load L_t (1) under a general proactive service strategy $\sigma := \{u_l(\tau)\}_{l,\tau}$, the time average expected cost minimization problem is thus written as

$$c^*(T) := \min_{\sigma} \limsup_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \mathbb{E}\left[C\left(L_t\right)\right] \tag{2}$$

$$\text{subject to, } \sum_{\tau=1}^{T} u_{l-\tau}(\tau) \le S, \quad \forall l = 0, 1, \cdots, \tag{3}$$

$$u_l(\tau) \ge 0, \quad \forall l = 0, 1, \cdots, \quad \tau = 1, \cdots, T. \tag{4}$$

The expectation operator $\mathbb{E}[.]$ is taken over all random variables in the system including all requests and proactive services. The constraints (3), (4) are to ensure that the SP can not proactively consume more than S units of resources to cache predictable demand for a future slot and that caching control is never negative.

The memory introduced to the system through correlated user activity *enhances* the predictability of future demand as the system gets closer to the actual demand instant. However, such correlated activity renders the problem more challenging

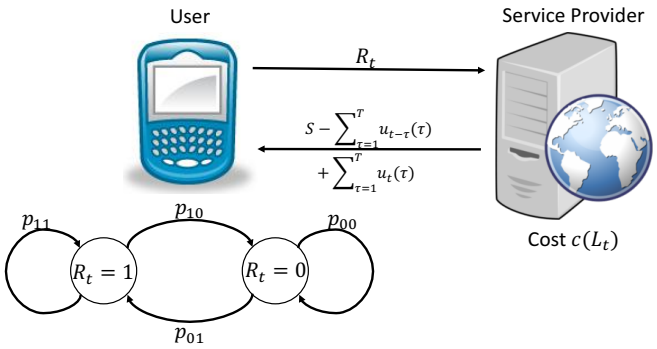


Fig. 1: User-SP interactions.

[1]The consumed resource S to serve an item need not be constant, it could be time dependent but the SP must know it before applying proactive service for that item.

to tackle compared to the previous work [7] where demand dynamics have been assumed independent every time slot.

III. Fundamental Lower Bound

In this section, we establish a fundamental lower bound on the achievable cost by any proactive service policy.

Theorem 1: For a given proactive service window T, the minimum achievable time average expected cost satisfies $c^*(T) \geq c_b(T)$, where

$$\begin{aligned} c_b(T) := \min_{\{\mu_\tau\}_{\tau=1}^T, \mu_\tau \geq 0} (1-\pi) C\left(\sum_{\tau=1}^T \mu_\tau\right) + \\ p_{11}\pi C\left(S - (1-\pi)\sum_{\tau=2}^T \mu_\tau (1-(1-(p_{10}+p_{01}))^{\tau-1})\right) + \\ p_{10}\pi C\left(S - \sum_{\tau=2}^T \mu_\tau (1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1\right) \quad (5) \\ \text{subject to, } \sum_{\tau=1}^T \mu_\tau \leq S, \end{aligned}$$

Proof. Please refer to Appendix A. ■

Theorem 1 yields a simple finite dimensional and convex optimization that serves as lower bound for any feasible caching policy. The optimization (5) has a unique solution for a given proactive service window T as the cost function C is strictly convex and the constraints set is compact. We denote such optimal solution by $\{\mu_\tau^*\}_\tau$.

Different from [7] where the lower bound has been characterized irrespective of the proactive window size, here we provide a tighter bound for every value of T. This will enable further insights on the capability of the proposed scheduling policy in regimes where T takes on moderate values.

In the following lemmas we highlight some key insights on the solution structure of (5) which are crucial to establish the merits of our proposed policy in Section IV.

Lemma 1: The optimal solution $\{\mu_\tau^*\}_{\tau=1}^T$ satisfies

$$\mu_1^* + \sum_{\tau=2}^{T} \mu_\tau^* (1-(p_{10}+p_{01}))^{\tau-1} \geq 0. \quad (6)$$

Proof. Please refer to Appendix B. ■

It is important to note from Lemma 1 that inequality (6) holds even when $p_{10} + p_{01} > 1$.

Lemma 2: For any $\tau \geq 2$, $\lim_{T\to\infty} \mu_\tau^* = 0$.

Proof. Please refer to Appendix C. ■

As the proactive service window T grows, there will be no particular preference over $\tau \in \{2, \cdots, T\}$ to maintain $\mu_\tau^* > 0$. In other words, the share of the optimal solution for every μ_τ^* diminishes asymptotically for $\tau \geq 2$.

In the following section, we present and analyze our proposed proactive caching policy.

IV. Proposed Proactive Caching Policy

Our proposed policy, denoted by $\bar{\sigma}$, is defined as follows.

Definition 1 (Policy $\bar{\sigma}$): Consider a service policy $\bar{\sigma}$ in which the proactive service control $u_t(\tau)$ is assigned as

$$u_t(\tau) := \begin{cases} \mu_1^*, & \tau = 1, R_t = 0, \\ \sum_{d=2}^T \frac{\mu_d^*}{T-1}, & \tau \in \{2, \cdots, T\}, R_t = 0, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where $\{\mu_\tau^*\}_{\tau=1}^T$ is the optimal solution of (5).

We can clearly see that our proposed policy is feasible as the cache controls $\{u_t(\tau)\}_t$ satisfy constraints (3), (4).

Before establishing its asymptotic optimality, we provide the following two remarks.

Remark 1: The policy $\bar{\sigma}$ aims at efficiently utilizing the temporally improving predictability and load balancing over time through its two main service components: (1) the near-demand service component (exploration) $u_t(1) = \mu_1^*$, which harnesses the most information about the demand of the next slot t, and (2) the load balancing component (exploitation) $\{u_t(\tau)\}_\tau = \frac{1}{T-1}\sum_{\tau=2}^T \mu_\tau^*$ which aims at smoothing out the load over time.

Remark 2: Policy $\bar{\sigma}$ is a simple stationary online policy. It assigns its controls based on *only* observing the current demand state R_t and *not* on the past.

With the Markov correlated nature of the demand process $\{R_t\}_t$, the SP gains more information about future demand as time gets closer to such demand instant. As such, the accuracy (or quality) of proactive caching improves over time thus attracting the SP to wait as close as possible to demand instant to apply proactive caching. On the other hand, the more the SP waits to apply a proactive caching service, the more it wastes opportunities for proactively caching content early on and spreading its service load uniform over more slots and reducing its operational cost.

Theorem 2 (Asymptotic optimality of $\bar{\sigma}$): Let $c_{\bar{\sigma}}(T)$ be the time average expected cost under the proposed policy $\bar{\sigma}$ where

$$\begin{aligned} c_{\bar{\sigma}}(T) = p_{11}\pi \mathbb{E}\Big[C\left(S - \sum_{\tau=2}^T u_{t-\tau}(\tau)\right) | R_t = 1, \\ R_{t-1} = 1\Big] + p_{10}\pi \mathbb{E}\Big[C\left(S - \sum_{\tau=2}^T u_{t-\tau}(\tau) - \mu_1^*\right) | \\ R_t = 1, R_{t-1} = 0\Big] + (1-\pi)\mathbb{E}\left[C\left(\mu_1^* + \sum_{\tau=2}^T u_t(\tau)\right)\right] \quad (8) \end{aligned}$$

Then $\lim_{T\to\infty} c_{\bar{\sigma}}(T) - c_b(T) = 0$.

Proof. Omitted for space limitation. However, it follows from Lemmas 1, 2. For complete proof, please refer to Appendix D in [12]. ■

Theorem 2 utilizes the fact that, as the proactive service window grows, the proactive scheduler observes the *typical* time average behavior of the Markov demand process where the impact of the past user activity on the future decisions diminish over time. Thus, proactive controls can be set equal, interestingly, up to two slots to the actual demand time while a more tailored control $u_t(1)$ is served one slot ahead.

Besides the asymptotic optimality property, policy $\bar{\sigma}$ is optimal for some non-asymptotic regimes that we discuss in the next section in addition to some other relevant scenarios.

V. Special Cases of the Proposed Policy

In this section, we study relevant network scenarios captured by our proposed correlated demand profile and investigate the performance of our proposed policy $\bar{\sigma}$.

Theorem 3: For any proactive service window $T \geq 1$, if $p_{10} = 1$, then $\mu_\tau^* = 0$, $\tau = 2, \cdots, T$ and policy $\bar{\sigma}$ is optimal.

Proof. Please refer to Appendix E in [12]. ■

If the user idles for at least one slot right after requesting a service (i.e., $p_{10} = 1$), then irrespective of the size of the proactive service window, the optimal proactive caching strategy is to defer the caching decision to only one slot before potential demand. Intuitively, in such a case, the SP is certain about experiencing at least one idle slot before demand. Thus it can safely utilize as much information as possible about future demand to maximally improve the accuracy of the proactive caching service while it is guaranteed to serve it in the slot before predicted demand. Note that, from the definition of policy $\bar{\sigma}$, the SP does not apply any proactive caching service in a slot experiencing user demand (i.e., when $R_t = 1$).

In the medium timescale (order of seconds), the $p_{10} = 1$ scenario is practically typical in applications dominated by user-SP interactions with the user spending an amount of time to process the content served by the SP before generating another content request. For instance, the study in [11] shows that end users interacting with Google Chrome and Yelp spend on average 3 and 2.5 seconds, respectively, processing previously requested content before sending a new request while the service of a request itself only takes, on average, 0.5 to 1 second. If the slot duration is 1 second, then the end user will likely idle more than one slot after every service.

Theorem 4: If the proactive service window size $T = 1$, the proposed policy $\bar{\sigma}$ is optimal. That is, $c_{\bar{\sigma}}(1) = c_b(1)$.

Proof. Follows by substituting with $T = 1$ in (5) and (8). ■

When the data content is highly dynamic, e.g., news, traffic, or some social network updates, the SP does not have enough freedom to proactively spread its future demand over time. As such, the near demand service component μ_1^* is the only degree of freedom to minimize the cost. The resulting cost in this case coincides with the lower bound implying the optimality of the policy $\bar{\sigma}$ irrespective of the user-SP interaction characteristics. That is, policy $\bar{\sigma}$ is optimal for any p_{10}, p_{01}.

The $T = 1$ scenario can also serve as a good approximation for any SP that can only make limited-term predictions of its user's data requests. In fact, we show numerically in Section VI that the *most contribution to cost reduction offered by proactive content caching is attained by the one-slot ahead service component* $u_t(1)$.

Theorem 5: If $p_{10} + p_{01} \leq 1$, then[2] $\mu_\tau^* = 0$ for all $\tau = 1, \cdots, T-1$, but $\mu_T^* > 0$.

Proof. Please refer to Appendix F in [12]. ■

If the user requests $\{R_t\}_t$ form an independent and identically distributed (i.i.d.) sequence rather than a Markov process, then $p_{10} + p_{01} = 1$, in which case Theorem 5 agrees with the asymptotic optimality result of Theorem 3 in [7] where users demand has been assumed i.i.d. The value of Theorem 5 is that it validates that the i.i.d. approximation is also asymptotically optimal for correlated demand scenarios with $p_{10} + p_{01} < 1$.

In general, $p_{10} + p_{01} < 1$ represents the scenario where the user tends to remain in one state (idle or request) more than to switch between states. Under such a condition, the SP does not have enough certainty about an empty slot prior to the demand in order to apply a near-demand proactive caching, hence it does not risk delaying proactive caching and starts early on distributing its load.

The $p_{10} + p_{01} < 1$ scenario can model the network's operation in the long and short timescales as follows. In the long timescale (order of minutes to hours), the user interacting with a video streaming SP may spend several minutes watching a group of movies in a row (during a break or after the business day). Then the user refrains from sending more movie requests (returning to work or sleeping) creating a sequence of idle slots, where the time slot itself can be the duration of consuming one video. It has been recently found that the average YouTube session duration is 40 minutes [13] while the average *popular* YouTube video length is 4.3 minutes [14] suggesting that users may watch more than one video in one session. Thus, the user may remain in the same state more than transitioning between states.

In the short timescale (order of milliseconds), the interactive-traffic study [11] has revealed that the response to a user query is a sequence of packets that occupy the transmission link for several milliseconds and possibly a whole second followed by a period of idling due to the user processing such response. The idle period also spans thousands of milliseconds. If the time slot is considered to be the packet transmission duration, then the packet transmission process of an interactive App can be modeled as a correlated Markov process with R_t being the event of packet transmission in slot t, and p_{10}, p_{01} are the corresponding transition probabilities.

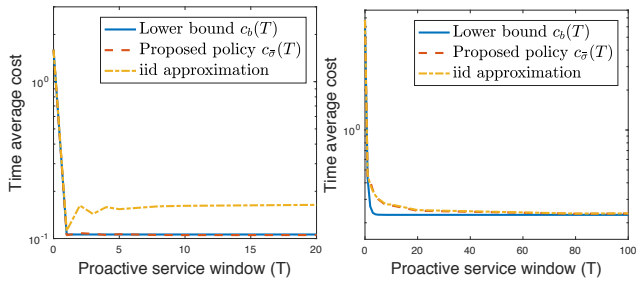
VI. Numerical Simulations

In this section, we provide numerical validation of the analytical results presented in the previous sections. In our simulations, we consider the i.i.d. approximation as our baseline caching strategy where the user's demand is assumed i.i.d. with probability of request $\pi = p_{01}$ in any time slot. The i.i.d. approximation ignores the additional information available to the SP due to demand correlation. We also consider a polynomial cost function of degree $d = 4$, i.e., $C(L) = L^4$, and set the service resource of a single request to $S = 1$ unit.

In Fig. 2, we show the time average cost performance as well as the fundamental lower bound's behavior with the proactive window size T for two instances of the user-SP interaction. First, $p_{10} + p_{01} > 1$ where $p_{10} = 0.9$, $p_{01} = 0.8$. Fig. 2a manifests that the cost under our proposed policy $\bar{\sigma}$ converges very fast in T, almost from $T = 1$. In addition, the figure demonstrates the *gap* between the cost under i.i.d. approximation and under policy $\bar{\sigma}$ which is $\sim 34\%$ of the i.i.d. approximation's cost. Second, $p_{10} + p_{01} < 1$ where $p_{10} = 0.3$, $p_{01} = 0.4$. Fig. 2b shows that the i.i.d. approximation's cost coincides with our proposed policy's validating Theorem 5. While convergence of the time average cost is slower than that of $p_{10} + p_{01} > 1$, it happens at reasonably moderate values of T. For instance, if the time slot duration in the long timescale is 10 minutes, then $T = 100$ corresponds to 16.67 hours, which can be considered a short lifetime for a wide range of video content like TV shows, songs, Vlogs, etc.

Overall, the Fig. 2 confirms Theorem 2 by showing that the cost under policy $\bar{\sigma}$ converges asymptotically to $c_b(T)$.

[2]There is no contradiction between Lemma 2 and Theorem 5 as Lemma 2 considers the limiting behavior of $\{\mu_\tau^*\}_{\tau=2}^T$ as $T \to \infty$.

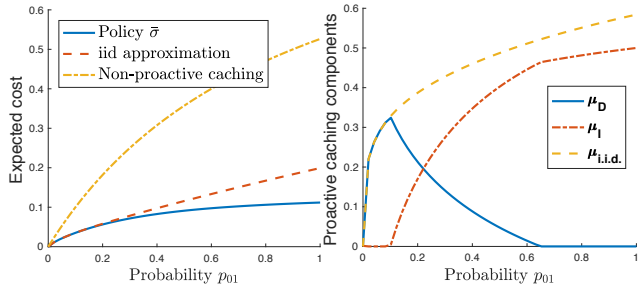


(a) $p_{10} = 0.9$, $p_{01} = 0.8$ (b) $p_{10} = 0.3$, $p_{01} = 0.4$

Fig. 2: Asymptotic optimality of the proposed policy $\bar{\sigma}$.

Moreover, it verifies Theorem 4 as in Figs. 2a, 2b we have $c_{\bar{\sigma}}(1) = c_b(1)$. We can also observe from the figures that transitioning from $T = 0$ (the reactive service scenario) to $T = 1$ leads to the most significant contribution to cost reduction, as mentioned in the discussion of Theorem 4 in Section V.

In Fig. 3, we shed light on the impact of the user *reaction* time, captured by p_{01} of interactive-app session on the performance of proactive caching. In particular, we set $p_{10} = 0.9$, and consider the medium timescale of operation where the time slot duration is a few seconds and the user is interacting with an app through a sequence of query-response pairs [11]. We assume that the served information has a lifetime much larger than the slot size, hence we take $T \to \infty$. We define the asymptotic service components of our policy $\bar{\sigma}$ as $\mu_I := \lim_{T\to\infty} \mu_1^*$, $\mu_D := \lim_{T\to\infty} \sum_{\tau=2}^{T} \mu_\tau^*$, where μ_I captures the near-demand service component, whereas μ_D captures the load-balancing service component.



(a) Cost vs p_{01} (b) Caching components vs p_{01}

Fig. 3: Asymptotic optimality of the proposed policy $\bar{\sigma}$.

Fig. 3a shows the time average cost grows with p_{01}. As the user reaction time lessens p_{01} grows adding more utilization to the SP's resources and hence raising the cost. The cost under i.i.d. approximation matches that of policy $\bar{\sigma}$ for small values of p_{01}. In particular, according to Theorem 5, for all $p_{01} \leq 1 - p_{10} = 0.1$, the i.i.d. approximation is optimal in our large T scenario. As p_{01} grows beyond 0.1, the i.i.d. approximation deviates from the optimal (policy $\bar{\sigma}$) while the cost of policy $\bar{\sigma}$ grows at slower rate highlighting the value of leveraging the near-demand component rather than simply distributing the load equally over the proactive service window.

Fig. 3b further depicts the allocation of the caching components as p_{01} grows. Typically, the system begins with $\mu_I = 0$ as $p_{01} + p_{10} \leq 1$, then as transitioning between states starts to dominate the system, i.e., $p_{10} + p_{01} > 1$, proactive caching starts to utilize the near demand component especially as the certainty about demand in the next slot, p_{01}, grows. In fact, it becomes more advantageous to the system to delay all its proactive service to only one slot before demand as the certainty, p_{01}, approaches one.

VII. Conclusion

We have considered optimal proactive content caching for an end user that exhibits a *correlated* demand pattern over time. Such correlation best captures realistic scenarios of today's network dynamics in various timescales including Apps that are based on user-service provider (SP) interactions like Yelp, Web-browsing, etc. The SP has been assumed to employ proactive caching within a proactive service window to best regulate its service load and minimize its operational cost. In this framework, we have developed an asymptotically optimal caching strategy that achieves a fundamental cost lower bound as the proactive service window grows. We have further characterized non-asymptotic cases where our proposed caching strategy remains optimal.

Appendix A
Proof of Theorem 1

Let $\{u_t^*(\tau)\}_{t,\tau}$ be the optimal solution to (2). Then, by conditioning on R_{l-1}, R_l, for $l \geq 0$ we have

$$
\begin{aligned}
c^*(T) &= \limsup_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \\
&p_{11}\pi\mathbb{E}\left[C\left(S - \sum_{\tau=1}^{T} u_{l-\tau}^*(\tau) + \sum_{\tau=1}^{T} u_l^*(\tau)\right) | R_{l-1} = 1, R_l = 1\right] + \\
&p_{10}\pi\mathbb{E}\left[C\left(S - \sum_{\tau=1}^{T} u_{l-\tau}^*(\tau) + \sum_{\tau=1}^{T} u_l^*(\tau)\right) | R_{l-1} = 0, R_l = 1\right] \\
&+ p_{01}(1-\pi)\mathbb{E}\left[C\left(\sum_{\tau=1}^{T} u_l^*(\tau)\middle| R_{l-1} = 1, R_l = 0\right]\right. \\
&+ p_{00}(1-\pi)\mathbb{E}\left[C\left(\sum_{\tau=1}^{T} u_l^*(\tau)\right) | R_{l-1} = 0, R_l = 0\right] \\
&\overset{(a)}{\geq} \limsup_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} p_{11}\pi C\left(S - \sum_{\tau=1}^{T} \mathbb{E}[u_{l-\tau}^*(\tau) \right. \\
&\left. | R_{l-1} = 1, R_l = 1] + \sum_{\tau=1}^{T} \mathbb{E}\left[u_l^*(\tau) | R_{l-1} = 1, R_l = 1\right]\right) +
\end{aligned}
$$

$$p_{10}\pi C\left(S-\sum_{\tau=1}^{T}\mathbb{E}\left[u^*_{l-\tau}(\tau)|R_{l-1}=0,R_l=1\right]+\sum_{\tau=1}^{T}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=0,R_l=1\right]\right)+p_{01}(1-\pi)C\left(\sum_{\tau=1}^{T}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=1,R_l=0\right]\right)+p_{00}(1-\pi)C\left(\sum_{\tau=1}^{T}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=0,R_l=0\right]\right)$$

$$\overset{(b)}{\geq}\limsup_{t\to\infty}p_{11}\pi C\left(S-\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-1}=1,R_l=1]+\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=1,R_l=1\right]\right)+p_{10}\pi C\left(S-\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\left[u^*_{l-\tau}(\tau)|R_{l-1}=0,R_l=1\right]+\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=0,R_l=1\right]\right)+p_{01}(1-\pi)C\left(\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=1,R_l=0\right]\right)+p_{00}(1-\pi)C\left(\sum_{\tau=1}^{T}\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}\left[u^*_l(\tau)|R_{l-1}=0,R_l=0\right]\right).$$

Inequality (a) follows by Jensen's inequality since C is strictly convex. Inequality (b) is another application of Jensen's inequality while noting that $\frac{1}{t}\sum_{l=0}^{t-1}1=1$.

The expectation $\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-1},R_l]$ can be expanded as

$$\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-1},R_l]=\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1},R_l)\times\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-\tau-1}=x,R_{l-\tau}=y,R_{l-1},R_l],\quad(9)$$

$$P_\tau(x,y|R_{l-1},R_l):=P(R_{l-\tau-1}=x,R_{l-\tau}=y|R_{l-1},R_l)\quad(10)$$

for any $x,y\in\{0,1\}$. The causality of the cache controls implies that the demand realization R_l is independent of the previous controls $\{u^*_{l-\tau}(\tau)\}$, $\tau=1,\cdots,T$. Hence, (9) can be re-written as

$$\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-1},R_l]=\sum_{x,y\in\{0,1\}}\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-\tau-1}=x,R_{l-\tau}=y]P_\tau(x,y|R_{l-1},R_l).$$

Let $\mu_\tau(xy):=\frac{1}{t}\sum_{l=T}^{t-T}\mathbb{E}[u^*_l(\tau)|R_{l-1}=x,R_l=y]$, $x,y\in\{0,1\}$. Then, we have

$$\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}[u^*_l(\tau)|R_{l-1}=x,R_l=y]\leq\mu_\tau(xy)+\epsilon_t,$$

$$\frac{1}{t}\sum_{l=0}^{t-1}\mathbb{E}[u^*_{l-\tau}(\tau)|R_{l-\tau-1}=x,R_{l-\tau}=y]\leq\mu_\tau(xy)+\epsilon_t,$$

where $\epsilon_t=\frac{TS}{t}$. Note that, $\epsilon_t\to0$ as $t\to\infty$.

From inequality (b) above, the optimal time average expected cost can further be lower bounded as

$$c^*(T)\geq\limsup_{t\to\infty}p_{11}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=1,R_l=1)\times(\mu_\tau(xy)+\epsilon_t)+\sum_{\tau=1}^{T}\mu_\tau(11)\right)+p_{00}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau(00)\right)+p_{10}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=0,R_l=1)\times(\mu_\tau(xy)+\epsilon_t)+\sum_{\tau=1}^{T}\mu_\tau(01)\right)+p_{01}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau(10)\right)$$

$$\overset{(c)}{\geq}\limsup_{t\to\infty}\min_{\{\mu_\tau(xy)\}_{\tau,x,y}}p_{01}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau(10)\right)+p_{11}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=1,R_l=1)\times(\mu_\tau(xy)+\epsilon_t)+\sum_{\tau=1}^{T}\mu_\tau(11)\right)+p_{10}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=0,R_l=1)\times(\mu_\tau(xy)+\epsilon_t)+\sum_{\tau=1}^{T}\mu_\tau(01)\right)+p_{00}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau(00)\right)$$

$$\text{subject to, }\mu_\tau(xy)\geq0,\quad x,y\in\{0,1\},\tau=1,\cdots,T$$

$$\sum_{\tau=1}^{T}\mu_\tau(xy)\leq S,\quad x,y\in\{0,1\}.$$

The optimization on the RHS of inequality (c) is strictly convex in $\{\mu_\tau(xy)\}_{\tau,x,y}$ and its constraint set is compact, hence it has a unique solution, call it $\{\mu^*_\tau(xy,\epsilon_t)\}_{\tau,x,y}$.

From Corollary 7.43 in [15], since C is strictly convex, then $\mu^*_\tau(xy,\epsilon_t)$ is continuous in ϵ_t, and $\mu^*_\tau(xy,\epsilon_t)\to\mu^*_\tau(xy)$ as $t\to\infty$. Hence, it follows that

$$c^*(T)\geq p_{11}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=1,R_l=1)\times\mu^*_\tau(xy)+\sum_{\tau=1}^{T}\mu^*_\tau(11)\right)+p_{00}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu^*_\tau(00)\right)+p_{10}\pi C\left(S-\sum_{\tau=1}^{T}\sum_{x,y\in\{0,1\}}P_\tau(x,y|R_{l-1}=0,R_l=1)\mu^*_\tau(xy)\sum_{\tau=1}^{T}\mu^*_\tau(01)\right)+p_{01}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu^*_\tau(10)\right).\quad(11)$$

Since C is monotonically increasing and $P_\tau\leq1$, the RHS of (11) can only increase in $\mu^*_\tau(x1)$, $x\in\{0,1\}$, $\tau=1,\cdots,T$.

Therefore, $\mu_\tau^*(11) = \mu_\tau^*(01) = 0$, $\tau = 1, \cdots, T$.

Expanding P_τ from (10) and applying Jensen's inequality to the RHS of (11) we obtain

$$c^*(T) \geq (1-\pi)C\left(\sum_{\tau=1}^{T}(p_{00}\mu_\tau^*(00) + p_{01}\mu_\tau^*(10))\right) + p_{11}\pi C\left(S - \sum_{\tau=2}^{T}(1-\pi)\left(1-(1-(p_{10}+p_{01}))^{\tau-1}\right)\times (p_{00}\mu_\tau^*(00) + p_{01}\mu_\tau^*(10))\right) + p_{10}\pi C\Bigg(S - \sum_{\tau=1}^{T}(1-\pi+\pi(1-(p_{10}+p_{01})^{\tau-1})(p_{00}\mu_\tau^*(00) + p_{01}\mu_\tau^*(10)))\Bigg) \tag{12}$$

Let $\mu_\tau := p_{00}\mu_\tau^*(00) + p_{01}\mu_{\tau(10)}$. Since $p_{00} + p_{10} = 1$, the optimal time average expected cost is lower bounded by

$$c^*(T) \geq \min_{\{\mu_\tau\}_{\tau=1}^T, \mu_\tau \geq 0, \sum_{\tau=1}^T \mu_\tau \leq S}(1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau\right) + p_{11}\pi C\left(S - (1-\pi)\sum_{\tau=2}^{T}\mu_\tau(1-(1-(p_{10}+p_{01}))^{\tau-1})\right) + p_{10}\pi C\left(S - \sum_{\tau=2}^{T}\mu_\tau(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1\right)$$

Appendix B
Proof of Lemma 1

When $p_{10}+p_{01} \leq 1$, the inequality (6) is trivial since $\mu_\tau^* \geq 0, \tau = 1, \cdots, T$. When $p_{10}+p_{01} > 1$, consider the Lagrangian of the optimization (5).

$$g(\lambda, \{\mu_\tau\}_\tau) := (1-\pi)C\left(\sum_{\tau=1}^{T}\mu_\tau\right) + \lambda\left(\sum_{\tau=1}^{T}\mu_\tau - S\right) + p_{11}\pi C\left(S - (1-\pi)\sum_{\tau=2}^{T}\mu_\tau(1-(1-(p_{10}+p_{01}))^{\tau-1})\right) + p_{10}\pi C\left(S - \sum_{\tau=2}^{T}\mu_\tau(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1\right). \tag{13}$$

The dual problem of (5) is, therefore,

$$\max_{\lambda \geq 0}\min_{\{\mu_\tau\}_{\tau=1}^T, \mu_\tau \geq 0} g(\lambda, \{\mu_\tau\}_\tau).$$

From the KKT conditions on the dual problem, we have: if $\mu_1^* > 0$, then $\frac{\partial g}{\partial \mu_1^*} = 0$. That is

$$\lambda + (1-\pi)C'\left(\sum_{\tau=1}^{T}\mu_\tau^*\right) - \pi p_{10}C'\Bigg(S - \sum_{\tau=2}^{T}\mu_\tau^* \times (1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1^*\Bigg) = 0. \tag{14}$$

Further, for any $\tau_0 \in \{2, \cdots, T\}$ such that $\mu_{\tau_0}^* > 0$, we have $\frac{\partial g}{\partial \mu_{\tau_0}^*} = 0$. That is:

$$\lambda + (1-\pi)C'\left(\sum_{\tau=1}^{T}\mu_\tau^*\right) - \pi p_{11}(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1})\right) - \pi p_{10}(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1^*\right) = 0. \tag{15}$$

Assuming that $\mu_1^* > 0$, substitute from (14) in (15) to get

$$\pi p_{10}(1-1+\pi-\pi(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1^*\right) = \pi p_{11}(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S - \sum_{\tau=2}^{T}\mu_{\tau_0}^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau_0-1})\right).$$

Noting that $\pi p_{10} = (1-\pi)p_{01}$, we obtain

$$p_{01}C'\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1^*\right) = p_{11}C'\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1})\right), \tag{16}$$

where C' is the first derivative of the cost function C.

Since C is convex, C' is monotonically increasing. Thus, its inverse exists. In addition, $p_{01} > p_{11}$ by the hypothesis that $p_{10}+p_{01} > 1$. Dividing (16) by p_{01} and taking C'^{-1} of both sides, we get

$$S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1}) - \mu_1^* = C'^{-1}\left(\frac{p_{11}}{p_{01}}C\left(S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1})\right)\right) \leq S - \sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1},$$

which implies that $\mu_1^* + \sum_{\tau=2}^{T}\mu_\tau^*(1-(p_{10}+p_{01}))^{\tau-1} \geq 0$.

Assuming that $\mu_1^* = 0$, then (14) does not hold. However, the gradient of the Lagrangian of $g(\lambda, \{\mu_\tau\}_\tau)$ at $\mu_\tau = \mu_\tau^*$ in

the direction of μ_1 can not be greater than its value in the direction of μ_{τ_0} for otherwise, μ_1^* must be positive. Thus,

$$p_{10}\pi C'\left(S-\sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1})\right)\leq \pi p_{11}(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S-\sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1})\right)+ \pi p_{10}(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau_0-1})\times C'\left(S-\sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1})\right),$$

$$\Rightarrow p_{01}C'\left(S-\sum_{\tau=2}^{T}\mu_\tau^*(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1})\right) \leq p_{11}C'\left(S-\sum_{\tau=2}^{T}\mu_\tau^*(1-\pi)(1-(1-(p_{10}+p_{01}))^{\tau-1})\right).$$

Again, dividing by p_{01}, taking the inverse C'^{-1} and noting that $p_{01} > p_{11}$, we get $\sum_{\tau=2}^{T}\mu_\tau^*(1-(p_{10}+p_{01}))^{\tau-1}\geq 0$ which completes the proof.

Appendix C
Proof of Lemma 2

It suffices to show that there exists a feasible solution $\{\hat{\mu}_\tau\}_{\tau=1}^{T}$ to (5) for which $\hat{\mu}_\tau=\frac{K}{T-1}+\epsilon^{\tau-1}$, for some $K>0$ and $\epsilon\in[0,1)$, $\tau=2,\cdots$, such that $|\mu_\tau^*-\hat{\mu}_\tau|\to 0$ as $T\to\infty$. To that end, $\{\hat{\mu}_\tau\}_\tau$ must satisfy the feasibility conditions: $\hat{\mu}_\tau\geq 0$ for all τ, $\sum_{\tau=1}^{T}\hat{\mu}_\tau\leq S$ and the optimality conditions:

$$\sum_{\tau=1}^{T}\hat{\mu}_\tau-\mu_\tau^*=0 \quad (17)$$

$$\sum_{\tau=2}^{T}(\hat{\mu}_\tau-\mu_\tau^*)(1-\pi)(1-(1-(p_{10}+p_{01})^{\tau-1})=0 \quad (18)$$

$$\sum_{\tau=2}(\hat{\mu}_\tau-\mu_1^*)(1-\pi+\pi(1-(p_{10}+p_{01}))^{\tau-1})= \mu_1^*-\hat{\mu}_1 \quad (19)$$

as $T\to\infty$. Subtracting (19) from (18),

$$\mu_1^*-\hat{\mu}_1=\frac{1-(p_{10}+p_{01})}{p_{10}+p_{01}}(1-(1-(p_{10}+p_{01}))^{T-1})\frac{K}{T-1}+ (1-(\epsilon(1-(p_{10}+p_{01})))^{T-1})\frac{\epsilon(1-(p_{10}+p_{01}))}{1-\epsilon(1-(p_{10}+p_{01}))}- \sum_{\tau=2}^{T}\mu_\tau^*(1-(p_{10}+p_{01}))^{\tau-1}. \quad (20)$$

From (17), $\hat{\mu}_1=\sum_{\tau=1}^{T}\mu_\tau^*-K-(1-\epsilon^{T-1})\frac{\epsilon}{1-\epsilon}$. Thus, substituting in (20), and letting $T\to\infty$, we obtain

$$K=\sum_{\tau=2}^{\infty}\mu_\tau^*(1-(1-(p_{10}+p_{01}))^{\tau-1})+ \epsilon\left(\frac{1-(p_{10}+p_{01})}{1-\epsilon(1-(p_{10}+p_{01}))}-\frac{1}{1-\epsilon}\right)$$

We next select a value for $\epsilon\in[0,1)$ to establish the feasibility conditions of $\{\hat{\mu}_\tau\}_\tau$. First, $\hat{\mu}_1\geq 0$, that is

$$\sum_{\tau=1}^{\infty}\mu_\tau^*-\sum_{\tau=2}^{\infty}\mu_\tau^*(1-(1-(p_{10}+p_{01}))^{\tau-1})- \epsilon\left(\frac{1-(p_{10}+p_{01})}{1-\epsilon(1-(p_{10}+p_{01}))}-\frac{1}{1-\epsilon}\right)-\frac{\epsilon}{1-\epsilon}\geq 0,$$

which implies

$$\frac{\epsilon(1-(p_{10}+p_{01}))}{1-\epsilon(1-(p_{10}+p_{01}))}\leq\mu_1^*+\sum_{\tau=2}^{T}\mu_\tau^*(1-(p_{10}+p_{01}))^{\tau-1}. \quad (21)$$

From Lemma 1, choosing $\epsilon=0$ satisfies (21). Second, for any $\tau\geq 2$, with $\epsilon=0$, $\hat{\mu}_\tau=K/(T-1)$, with $K=\sum_{\tau=2}^{\infty}\mu_\tau^*(1-(1-(p_{10}+p_{01}))^{\tau-1})\geq 0$ which ensures that $\hat{\mu}_\tau\geq 0$.

References

[1] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains," *Information Theory, IEEE Transactions on*, vol. 59, pp. 4833–4854, Aug 2013.

[2] S. Zhang, L. Huang, M. Chen, and X. Liu, "Proactive serving decreases user delay exponentially: The light-tailed service time case," *IEEE/ACM Trans. Netw.*, vol. 25, pp. 708–723, Apr. 2017.

[3] J. Tadrous, H. El Gamal, and A. Eryilmaz, "Can carriers make more profit while users save money?," *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 1757–1761, June 2014.

[4] F. Alotaibi, S. Hosny, H. E. Gamal, and A. Eryilmaz, "A game theoretic approach to content trading in proactive wireless networks," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2216–2220, June 2015.

[5] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Joint smart pricing and proactive content caching for mobile services," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 2357–2371, Aug 2016.

[6] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Transactions on Networking*, vol. 23, pp. 1917–1930, Dec 2015.

[7] J. Tadrous and A. Eryilmaz, "On optimal proactive caching for mobile networks with demand uncertainties," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 2715–2727, October 2016.

[8] L. Huang, S. Zhang, M. Chen, and X. Liu, "When backpressure meets predictive scheduling," *The 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 33–42, 2014.

[9] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, pp. 82–89, Aug 2014.

[10] R. A. Hassan, A. M. Mohamed, J. Tadrous, M. Nafie, T. ElBatt, and F. Digham, "Dynamic proactive caching in relay networks," in *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–8, May 2017.

[11] J. Tadrous and A. Sabharwal, "Interactive app traffic: An action-based model and data-driven analysis," in *2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–8, May 2016.

[12] "On optimal proactive caching with correlated predictions - technical report." https://www.dropbox.com/s/gcdxs7zlctddwuo/Tech_report_Allerton2018.pdf?dl=0.

[13] C. D. Looper, "People now spend an average of 40 minutes on youtube per viewing session." http://www.techtimes.com/articles/69850/20150717/people-now-spend-average-40-minutes-youtube-per-viewing-session.htm, 2015.

[14] Minimatters, "The best video length for different videos on youtube." https://www.minimatters.com/youtube-best-video-length/, 2017.

[15] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*. Springer, 2009.