Learning-Based Content Caching with Time-Varying Popularity Profiles

B. N. Bharath, K. G. Nagananda, D. Gündüz, and H. Vincent Poor

***Abstract*—Content caching at the small-cell base stations (sBSs) in a heterogeneous wireless network is considered. A cost function is proposed that captures the backhaul link load called the "offloading loss", which measures the fraction of the requested files that are not available in the sBS caches. Previous approaches minimize this offloading loss assuming that the popularity profile of the content is time-invariant and perfectly known. However, in many practical applications, the popularity profile is unknown and time-varying. Therefore, the analysis of caching with *non-stationary* and statistically *dependent* popularity profiles (assumed unknown, and hence, estimated) is studied in this paper from a learning-theoretic perspective. A probably approximately correct (PAC) result is derived, in which a high probability bound on the offloading loss difference, *i.e.*, the error between the estimated (outdated) and the optimal offloading loss, is investigated. The difference is a function of the Rademacher complexity of the set of all probability measures on the set of cached content items, the $\beta-$mixing coefficient, $1/\sqrt{t}$ (t is the number of time slots), and a measure of discrepancy between the estimated and true popularity profiles.**

I. Introduction

A potential drawback of the small-cell infrastructure to offload wireless data from a macro base station (BS) is that the backhaul link-capacity required to support the peak data traffic can be extremely high, necessitating complex and expensive solutions to ensure high throughput and performance during peak traffic periods. Caching can reduce the peak load by shifting part of the traffic to off-peak hours by storing popular content items in the cache memories located at small-cell base stations (sBSs) during off-peak traffic periods [1]. The benefit of coded caching across sBSs is shown in [2], while in [3] caching is analyzed for networks modeled spatially using independent Poisson point processes (PPPs). In [4], proactive caching is shown to increase energy efficiency.

Most prior work in this area, including [2] - [4], assumes *a priori* knowledge of the popularity profile of the cached content, which is unreasonable in practical scenarios. This assumption is relaxed in [5] - [8], and various learning-based approaches are proposed to estimate the popularity profile, while a theoretical analysis has been carried out in [9] to study the implications of learning the popularity profile on the performance. However, these works assume that the popularity profile is fixed, and the requests are assumed to be independent across time. In practice, there are many applications (for example, video on demand) in which the popularity profile of cached content is time-varying [10]. Motivated by the growing significance of caching in improving the quality of service for end-users during peak traffic periods, we analyze the performance of a random caching strategy for a *non-stationary* popularity profile, which may exihibit statistical dependence across time.

A heterogenous network in which the users, BSs and sBSs, are distributed according to independent PPPs is considered. The sBSs employ a random caching strategy. A protocol model for communication is proposed, and a cost function, which captures the backhaul link overhead called the "offloading loss", is considered. The offloading loss at time t, which depends on the popularity profile, is denoted by $\mathcal{T}(t)$. Our goal is to obtain risk bounds on this offloading loss when the popularity profile is time-varying and unknown. Assuming a request model (see `Assumption 1`), the BS first estimates the popularity profile based on the requests observed during the first t slots. It then chooses the caching probabilities $\pi \triangleq (\pi_1, \pi_2, \dots, \pi_N)$, where N is the number of popular content items that can be cached, in order to minimize its offloading loss $\hat{\mathcal{T}}(t)$, based on the estimated popularity profile. sBSs in the coverage area of the BS use this optimal caching policy to store content items in their caches. Since the popularity profile is time-varying, it becomes necessary to frequently refresh the caches, say after every T time slots, albeit at an additional cost. Thus, it is important to investigate the minimum periodicity T of cache updates that guarantees the desired offloading loss.

In this paper, we derive probably approximately correct (PAC) type guarantees on the offloading loss difference $\Delta_{\mathcal{T}}(t, T)$, which is defined as the difference between the offloading loss incurred by using the outdated caching policy obtained by optimizing $\hat{\mathcal{T}}(t)$ at time $t + T$, and the optimal offloading loss at time $t + T$. We show that $\Delta_{\mathcal{T}}(t, T) < \epsilon$ with a probability of at least $1 - \delta$ for any $\delta > \zeta$ and $\epsilon > 0$, where ζ is a function of the β-mixing coefficient, the number of content items N comprising the popularity profile and the user density. The β-mixing coefficient is a measure of the statistical dependency of the time-varying popularity profiles. If the popularity profile process is "sufficiently" mixing, *i.e.*, if the process becomes almost independent after a sufficiently long time, and if the user density is very high, then the desired ϵ can be achieved for negligibly small $\delta > 0$. In particular, to

B. N. Bharath is with PES Institute of Technology, Bangalore South Campus, INDIA, E-mail: bharathbn@pes.edu. K. G. Nagananda is with PES University, INDIA, E-mail: kgnagananda@pes.edu. D. Gündüz is with Imperial College London, UK, E-mail: d.gunduz@imperial.ac.uk. H. V. Poor is with Princeton University, Princeton, NJ, USA, E-mail: poor@princeton.edu. This work was supported in part by the U.S. National Science Foundation under Grants CCF-1420575, DST/INT/UK/P-129/2016 and CNS-1456793.

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

achieve a fixed probability $\delta > \zeta$, we require the error ϵ to be a function of N, the rate of change of the popularity profile, and the Rademacher complexity, which is a measure of the difficulty in estimating the offloading loss.

The following are the main findings of this paper: (1) the error ϵ increases with N; (2) the desired error ϵ can be achieved with higher probability (i.e., ζ becomes smaller) for a larger user density, thus improving the caching performance, since higher user density results in more user-requests, allowing a better estimate of the popularity profile; (3) the higher the correlation of the popularity profile across time (defined in terms of the $\beta-$mixing coefficient), the longer the waiting time t to achieve a target error level ϵ with probability $1-\delta$; (4) the error ϵ is a function of the rate of change of the popularity profile, and hence T. Thus, outdated cache content results in a larger error for a given δ, and a rapidly varying popularity profile requires more frequent updates to achieve the desired error performance; (5) a higher Rademacher complexity results in poorer error performance; and (6) when the user requests are independent and identically distributed (i.i.d.), the error performance is better compared to non-stationary and statistically dependent requests. For stationary popularity profiles and large t, frequent cache-updates are not necessary to achieve the desired performance. To the best of our knowledge, this is the first time random caching is studied with non-stationary, statistically dependent and unknown popularity profiles from the standpoint of learning theory.

II. System Model

A heterogenous cellular network is considered in which the users, BSs and sBSs are spatially distributed according to independent PPPs with densities λ_u, λ_b and λ_s, respectively [11]. The sets of users, BSs and sBSs are denoted by $\Phi_u \subseteq \mathbb{R}^2$, $\Phi_b \subseteq \mathbb{R}^2$, and $\Phi_s \subseteq \mathbb{R}^2$, respectively. Each user requests a content item (or *files*) from the library $\mathcal{F} \triangleq \{f_1, \ldots, f_N\}$ of N files, each of size B bits, from its neighboring sBSs. The requests are assumed to be statistically independent across users. However, the requests from each user are assumed to be *non-stationary* and statistically *dependent* across time. We assume that the size of the cache at each sBS is at most M files. The problem considered in this paper is that of caching relevant "popular" files at the sBSs, wherein, depending on the availability of the file in the local cache, the requested file from the user will be served directly by one of the neighboring sBSs. In order to access cached content items, a user $u \in \Phi_u$ identifies and communicates with a set of neighboring sBSs employing the following protocol: Each sBS s located at $x_s \in \Phi_s$ communicates with a user u located at $x_u \in \Phi_u$ if $\|x_u - x_s\| < \gamma$, for some $\gamma > 0$. This condition determines the communication radius. In this protocol, we ignore the interference from other users in the network. The set of potential neighbors of user u located at x_u is denoted by $\mathcal{N}_u \triangleq \{y \in \Phi_s : \|y - x_u\| < \gamma\}$. The caching policy will depend on the distribution of the requests from the users, which is assumed to be unknown, and will be estimated. In the next subsection, we present a stochastic process modeling the requests from the users, and devise a method for estimating its distribution.

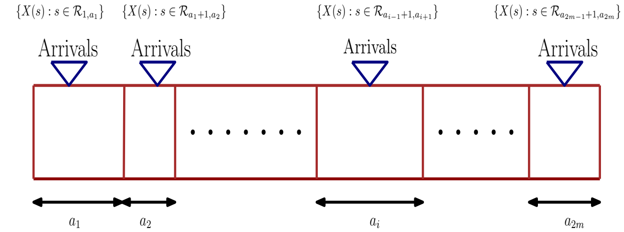

Fig. 1: This figure shows a time period consisting of t time slots each of duration Δ, which is divided into $2m$ blocks, where the i-th block is of size a_i slots, and $t = \sum_{i=1}^{2m} a_i$.

A. User Request Model

Let the stochastic process $X_v(\tau) \in \{1, 2, \ldots, N\}$ denote the index corresponding to a request from a user $v \in \Phi_u$ at time $\tau \in \mathbb{R}$. For example, each user can maintain an independent local Poisson clock, and makes a request whenever the local clock ticks. For any two users $v, w \in \Phi_u$, the request processes $X_v(\tau)$ and $X_w(\tau)$ are independent. For the ease of analysis, let us divide the time into slots of size $\Delta > 0$ each. Further, for each $v \in \Phi_u$, $\{X_v(\tau), \tau \in \mathbb{R}\}$ is a non-stationary and statistically dependent stochastic process across time slots, but the process $X_v(\tau)$ within each time slot (i.e., $\tau \in [i\Delta, (i+1)\Delta)$, $i = 1, 2, \ldots$) is assumed to be stationary. Further, we assume that there is a "typical" BS at the origin with a coverage radius of $R > 0$. The BS estimates the popularity of the content items based on the requests it receives. Essentially, at a given time slot t, the BS collects requests (for t time slots) from all the users in the BS's coverage area to estimate the popularity profile of the requested files. Let $n_u \sim \text{Poiss}(\pi\lambda_u R^2)$ denote the number of users in its coverage area. The arrival instants of the requests from different users are assumed to be random, and their distribution satisfies the following assumption.

`Assumption 1:` There exist constants $0 \leq \alpha_{\texttt{min}} \leq \alpha_{\texttt{max}} \leq 1$ such that for any random $n \geq 1$ users in the coverage area of the BS, the number of requests in $a \in \mathbb{N}$ time slots, denoted by $r_a \in \mathbb{N}$, satisfies $\alpha_{\texttt{min}} na \leq r_a \leq \alpha_{\texttt{max}} na$.

It turns out that the results based on the above assumption can be used to derive performance guarantees when the arrival process is a homogenous Poisson point process. Further, we assume that the request instants and the number of requests within a time slot are independent of the files requested. The set of request instants at which the requests from all the users in the coverage area of the BS arrive within the i^{th} time slot is denoted by $\mathcal{R}_i$. Let $X(s) \triangleq \bigcup_{v \in \Phi_u \bigcap \|X_v\|_2 \leq R} \{X_v(s)\}$ denote the set of requests from all the users in the coverage area of the BS at time $s \in \mathbb{R}$. The set of requests from all the users in time slots t_1 to t_2 is denoted by $X_{t_1,t_2} \triangleq \{X(s) : s \in \mathcal{R}_{t_1,t_2}\}$, where $\mathcal{R}_{t_1,t_2} \triangleq \bigcup_{i=t_1}^{t_2} \mathcal{R}_i$ (see Fig. 1). After receiving requests $X_{1,t}$ within first t time slots, the BS computes the empirical estimate of the popularity profile, *i.e.*, the probability of the i^{th} file being requested in time slot t, as follows:

$$\hat{p}_{i,t} = \frac{1}{r_t} \sum_{s \in \mathcal{R}_{1,t}} \mathbf{1}\{X(s) = i\},\ i = 1, \ldots, N, \tag{1}$$

where $r_t \triangleq |\mathcal{R}_{1,t}|$ is the total number of requests in the first t slots. The accuracy of the estimate $\hat{\mathcal{P}}^{(t)} \triangleq \{\hat{p}_{i,t} : i = 1, 2, \ldots, N\}$ depends on (i) the number of available samples, which in turn is related to the number of users in the coverage area of the BS, (ii) the number of requests per user, and (iii) the behavior of the process $X(s)$. The estimate in (1) is valid only when there is a positive number of user requests, which is guaranteed by `Assumption 1` above. In the next section, we present the performance measure for the above model, and state the main problem addressed in the paper.

III. Problem Statement

We consider a typical user located at the origin denoted by $o \in \Phi_u$. At time slot $t \in \mathbb{N}$, the "offloading loss" is defined as

$$\mathcal{T}(\Pi^{(t)}, \mathcal{P}^{(t)}, X_{1,t-1}) \triangleq \frac{B}{R_0} \Pr\left\{f_o \notin \mathcal{N}_u \mid X_{1,t-1}\right\}, \tag{2}$$

where $\Pi^{(t)}$ denotes the caching policy, $\mathcal{P}^{(t)} \triangleq \{p_1(t), p_2(t), \ldots, p_N(t)$ is the popularity profile in slot t, R_0 and $\frac{B}{R_0}$ denote the rate supported by the BS and the time overhead incurred in transmitting the file from the BS to the user, respectively, and f_o denotes the file requested by the typical user in the t-th slot. In the above, with a slight abuse of notation, $f_0 \notin \mathcal{N}_u$ is used to denote the event that the requested file f_0 is not present in the caches of the neighboring sBSs. The offloading loss is the scaled probability of the content requested by user o not being cached by any of the sBSs within its communication range conditioned on the requests received by the BS until the beginning of time slot t, *i.e.*, $X_{1,t-1}$. We employ the following random caching strategy, which enables us to derive a closed form expression for the offloading loss at time t.

Random Caching strategy: At time t (determined by the BS), each sBS $s \in \Phi_s$ caches content items in an i.i.d. fashion by generating M indices distributed according to $\Pi^{(t)} \triangleq \left\{\pi_i(t) : \sum_{i=1}^{N} \pi_i(t) = 1, \right\}$ (see [12]).

We seek to solve the following optimization problem:

$$\min_{\Pi^{(\tau)} \in P_\pi : \tau \in N} \quad \limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathcal{T}(\Pi^{(\tau)}, \mathcal{P}^{(\tau)}, X_{1,\tau-1}), \tag{3}$$

where $\mathcal{P}_\pi$ denotes the $N-$dimensional probability simplex. An expression for $\mathcal{T}(\Pi^{(t)}, \mathcal{P}^{(t)}, X_{1,t-1})$ is given in the following theorem whose proof can obtained by replacing p_i by $p_{X,i}(t)$ in the proof of Theorem 1 found in [9, Appendix A].

Theorem 1: The average offloading loss at time t for the random caching strategy $\Pi^{(t)}$ is given by

$$\mathcal{T}(\Pi^{(t)}, \mathcal{P}^{(t)}, X_{1,t-1}) = \left[\sum_{i=1}^{N} g(\pi_i(t)) p_{X,i}(t)\right], \tag{4}$$

where $p_{X,i}(t) \triangleq \Pr\{f_i \text{ requested by } o \text{ in slot } t | X_{1,t-1}\}$, and $g(\pi_i(t)) \triangleq \frac{B}{R_0} \exp\{-\lambda_u \pi \gamma^2 [1 - (1 - \pi_i(t))^M]\}$.

Even assuming that the conditional probabilities $p_{X,i}(t)$ are perfectly known, the complexity involved in solving (3) can be high owing to the fact that the caching policy at time t depends on $X_{1,t}$, which grows with t. In practice, the conditional probability $\Pr\{f_i \text{ requested } | X_{1,t-1}\}$ is unknown, and has to be estimated. Also, the BS may not have enough samples to compute a reasonably good estimate of this conditional probability. Hence, it is reasonable to consider the unconditional probability in the definition of the offloading loss. Thus, one can minimize the offloading loss $\mathcal{T}(\Pi^{(t)}, \mathcal{P}^{(t)}) \triangleq \left[\sum_{i=1}^{N} g(\pi_i(t)) p_i(t)\right]$, where $p_i(t)$ is the probability of the i^{th} file being requested at time t. However, the $p_i(t)$'s are unknown; and hence, an estimate of the popularity profile needs to be used in place of $\mathcal{P}^{(t)}$. More precisely, at time t, let $\hat{\Pi}_t^*$ denote the caching policy obtained using an estimate $\hat{\mathcal{P}}^{(t)}$, *i.e.*,

$$\hat{\Pi}_t^* = \arg\min_{\Pi^{(t)} \in \mathcal{P}_\pi} \quad \mathcal{T}(\Pi^{(t)}, \hat{\mathcal{P}}^{(t)}). \tag{5}$$

Suppose the cache content items chosen by the optimal caching policy at time t will be used to satisfy user demands over the period $(t, t+T]$. Let us consider the offloading loss in using $\hat{\Pi}_t^*$ at a later time, say at time $t+T$. The offloading loss at time $t+T$ is given by $\hat{\mathcal{T}}^*(t+T) \triangleq \mathcal{T}(\hat{\Pi}_t^*, \mathcal{P}^{(t+T)})$. Further, let Π_{t+T}^* denote the optimal caching policy at time $t+T$ using perfect knowledge of the popularity profile $\mathcal{P}^{(t+T)}$, *i.e.*,

$$\Pi_{t+T}^* = \arg\min_{\Pi^{(t+T)} \in \mathcal{P}_\pi} \quad \mathcal{T}(\Pi^{(t+T)}, \mathcal{P}^{(t+T)}), \tag{6}$$

with the corresponding offloading loss $\mathcal{T}^*(t+T) \triangleq \mathcal{T}(\Pi_{t+T}^*, \mathcal{P}^{(t+T)})$. Similar to [9], the central theme of this paper is the analysis of the offloading loss gap $\Delta_{\mathcal{T}}(t, T) \triangleq \hat{\mathcal{T}}^*(t+T) - \mathcal{T}^*(t+T)$. For example, if $\Delta_{\mathcal{T}}(t, T)$ is small, then each term in (3) is small, which results in a small average offloading loss. Moreover, this approach is central to analyzing the prediction problems involving non-stationary stochastic processes [13]. Next, we present the main result of this paper.

IV. Main Results

We study risk bounds on the offloading loss difference, $\Delta_{\mathcal{T}}(t, T)$ when the popularity profile is non-stationary. Essentially, for any $\epsilon > 0$, we seek to identify a risk bound $\delta > 0$, such that

$$\Pr\left\{\hat{\mathcal{T}}^*(t+T) - \mathcal{T}^*(t+T) > \epsilon\right\} < \delta. \tag{7}$$

First, we relate (7) to an expression in terms of the estimation error in the following theorem.

Theorem 2: For the estimate of the popularity profile in (1), the following bound holds:

$$\Pr\left\{\hat{\mathcal{T}}^*(t+T) - \mathcal{T}^*(t+T) > \epsilon\right\} \leq 2 \Pr\left\{\mathcal{A}_T(X_{1,t}) > \epsilon\right\},$$

where $\mathcal{A}_T(X_{1,t}) \triangleq \sup_{\Pi \in \mathcal{P}_\pi} \left|\sum_{i=1}^{N} g(\pi_i)(\hat{p}_{i,t} - p_{i,t+T})\right|$, and $g(\pi_i)$ is defined in Theorem 1.

Proof See [14, Appendix A].

The term $\Pr\{\mathcal{A}_T(X_{1,t}) > \epsilon\}$ can be bounded as follows:

$$\begin{aligned}\Pr\{\mathcal{A}_T(X_{1,t}) > \epsilon\} &= \sum_{j=0}^{\infty} \alpha_{t,T,\epsilon}^{(u)}(j) \\ &\leq \Pr\{n_u = 0\} + \sum_{j=1}^{\infty} \alpha_{t,T,\epsilon}^{(u)}(j) \\ &= \exp\{-\lambda_u \pi R^2\} + \sum_{j=1}^{\infty} \alpha_{t,T,\epsilon}^{(u)}(j),\end{aligned}$$

where $\alpha_{t,T,\epsilon}^{(u)}(j) \triangleq \Pr\{\mathcal{A}_T(X_{1,t}) > \epsilon \mid n_u = j\} \Pr\{n_u = j\}$. We next derive an upper bound on $\Pr\{\mathcal{A}_T(X_{1,t}) > \epsilon | n_u = j\}$. The term $\mathcal{A}_T(X_{1,t})$ depends on $\hat{p}_{i,t}$, which involves the sum of non-stationary random variables (RVs) which are possibly correlated across time. In order to apply the standard large deviation bounds, we must convert the sum of non-stationary dependent RVs to a sum of blocks of independent random vectors through a coupling argument, which is explained later in this section. For a given stochastic process $X_{1,\infty}$, and $s \in \mathbb{N}$, let $\mathbb{P}_{\tau,\tau+s}(\star)$ and $\mathbb{P}_{1\to\tau}(\star \mid \mathcal{E}) \otimes \mathbb{P}_{\tau+s\to\infty}(\star)$ denote the joint and product distributions of the stochastic processes $X_{1,\tau}$ and $X_{\tau+s,\infty}$, respectively. If $X_{1,\tau}$ and $X_{\tau+s,\infty}$ are independent, then $\|\mathbb{P}_{\tau,\tau+s}(\star) - \mathbb{P}_{1\to\tau}(\star) \otimes \mathbb{P}_{\tau+s\to\infty}(\star)\|_{\text{TV}} = 0$. Thus, for a given s, this difference, maximized over all $1 \leq \tau \leq \infty$ is a natural measure of the dependency between $X_{1,\tau}$ and $X_{\tau+s,\infty}$. This is commonly referred to as the $\beta-$mixing coefficient. For $s \in \mathbb{N}$, the β-mixing coefficient is given by

$$\beta(s) \triangleq \sup_{1\leq\tau\leq\infty} \|\mathbb{P}_{\tau,\tau+s}(\star) - \mathbb{P}_{1\to\tau}(\star) \otimes \mathbb{P}_{\tau+s\to\infty}(\star)\|_{TV}. \quad (8)$$

A stochastic process is said to be β-mixing if $\beta(s) \to 0$ as $s \to \infty$. For a given stochastic process that is β-mixing, two well-separated sequences of the process are approximately independent, where the approximation error is given by $\beta(s)$. Thus, we assume that the request process $X(t)$ is a β-mixing stochastic process, i.e., $\beta(s) \to 0$ as $s \to \infty$.

We now provide the details of the coupling argument, through which the dependent stochastic process is replaced by independent blocks of random variables. This will facilitate the use of a concentration inequality; in particular, McDiarmid's inequality. Fix $m \in \mathbb{N}$, and consider a sequence of consecutive blocks of size $a_i \in \mathbb{N}$, $i = 1, 2, \ldots, 2m$, slots such that $t = \sum_{j=1}^{2m} a_j$ (see Fig. 1). Let $a_0 = 0$. Consider the time instants at which the requests arrive corresponding to odd and even blocks defined as $\mathbb{T}_o^{(t)} \triangleq \bigcup_{j:j=0,2,4,\ldots,2(m-1)} \mathcal{R}_{a_j+1,a_{j+1}}$ and $\mathbb{T}_e^{(t)} \triangleq \bigcup_{j:j=1,3,5,\ldots,2m-1} \mathcal{R}_{a_j+1,a_{j+1}}$, respectively. Thus, the requests corresponding to the odd and even blocks are given by $X_{1,t}^e \triangleq \{X(s) : s \in \mathbb{T}_e^{(t)}\}$ and $X_{1,t}^o \triangleq \{X(s) : s \in \mathbb{T}_o^{(t)}\}$, respectively. In order to use a coupling argument, for a fixed $\mathcal{R}_{1,t}$, we define a new stochastic processes $\tilde{X}_{1,t}^h \triangleq \{\tilde{X}(s) : s \in \mathbb{T}_h^{(t)}\}$, $h \in \{e, o\}$, such that the requests in the even (and odd) blocks of $\tilde{X}_{1,t}$ are independent. However, within each block, the RVs can be arbitrarily correlated. Further, $\{\tilde{X}(s) : s \in \mathcal{R}_{a_{i-1}+1,a_i}\}$ and $\{X(s) : s \in \mathcal{R}_{a_{i-1}+1,a_i}\}$ have the same distribution, $i = 1, 2, \ldots, 2m$. We can always construct such a stochastic process, and the pair $(X(s), \tilde{X}(s))$ is called a *coupling* (see Fig. 1). We define $\tilde{X}_{1,t}^e$ and $\tilde{X}_{1,t}^o$ similarly to $X_{1,t}^e$ and $X_{1,t}^o$. The following theorem provides a bound on the performance guarantees in terms of the $\beta-$mixing coefficient.

Theorem 3: For the given model, and the popularity estimate in (1), with a probability of at least $1 - \delta$, $\delta > 2(\exp\{-\lambda_u \pi R^2\} + \sum_{i=2}^{2m-1} \beta(a_i))$, the following holds[1]:

$$\begin{aligned}\hat{\mathcal{T}}^*(t+T) - \mathcal{T}^*(t+T) &< \min\{\mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^e)], \mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^o)]\} \\ &+ \frac{N\alpha_{\max} B a_{\max}}{\alpha_{\min} R_0 a_{\min}} \sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{2m}}. \quad (9)\end{aligned}$$

In (9), $\delta' \triangleq \delta/2 - \exp\{-\lambda_u \pi R^2\} - \sum_{i=2}^{2m-1} \beta(a_i) > 0$, and

$$\mathcal{A}_T(\tilde{X}_{1,t}^{(h)}) \triangleq \sup_{\Pi \in \mathcal{P}_\pi} \left| \sum_{i=1}^{N} g(\pi_i) \left(\hat{p}_{i,t}^h - p_{i,t+T}\right) \right|, \quad (10)$$

where $\hat{p}_{i,t}^h \triangleq \frac{1}{\left|\mathbb{T}_h^{(t)}\right|} \sum_{s\in\mathbb{T}_h^{(t)}} \mathbf{1}\{\tilde{X}(s) = i\}$, $h \in \{e, o\}$.

Proof See Appendix A.

Next, we bound $\min\{\mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^e)], \mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^o)]\}$ to get the desired result. The bound that we derive depends on the Rademacher complexity and the nonstationarity of the stochastic process. We begin with the following definition.

Definition 1: **(Rademacher complexity)** The Rademacher complexity of $\mathcal{P}_\pi$ is defined as [15, Chapter 3]

$$\mathcal{R}_h^{(t)} \triangleq \mathbb{E}_{\tilde{X},\boldsymbol{\sigma}} \frac{1}{\left|\mathbb{T}_h^{(t)}\right|} \sup_{\Pi\in\mathcal{P}_\pi} \sum_{i=1}^{N} g(\pi_i) \Big| \sum_{s\in\mathbb{T}_h^{(t)}} \sigma_{i,s} \mathbf{1}\{\tilde{X}(s) = i\} \Big|,$$

where the Rademacher RVs $\sigma_{i,s} \in \{-1, 1\}$, $i = 1, 2, \ldots, N$ and $s \in \mathbb{T}_h^{(t)}$ are i.i.d. with probability $1/2$, $\boldsymbol{\sigma} \triangleq \{\sigma_{i,s} \in \{-1, 1\} : i = 1, 2, \ldots, N, s \in \mathbb{T}_h^{(t)}\}$, and $h \in \{e, o\}$.

Next, we provide the main result of this paper.

Theorem 4: For the given model and the popularity estimate in (1), with a probability of at least $1 - \delta$, $\delta > 2(\exp\{-\lambda_u \pi R^2\} + \sum_{i=2}^{2m-1} \beta(a_i) > 0)$, the following holds:

$$\begin{aligned}\hat{\mathcal{T}}^*(t+T) &< \mathcal{T}^*(t+T) + \max\{\mathcal{R}_e^{(t)}, \mathcal{R}_o^{(t)}\} \\ &+ \max\{\Delta_{t,T}^{(e)}, \Delta_{t,T}^{(o)}\} + \frac{N\alpha_{\max} B a_{\max}}{R_0 a_{\min} \alpha_{\min}} \sqrt{\frac{a_{\max} \log\left(\frac{1}{\delta'}\right)}{t}}, \quad (11)\end{aligned}$$

where $\mathcal{R}_h^{(t)}$ is the Rademacher complexity, $a_{\max} \triangleq \max_{1\leq i\leq 2m} a_i$, $\Delta_{t,T}^{(h)} \triangleq \sup_{\Pi\in\mathcal{P}_\pi} \sum_{i=1}^{N} g(\pi_i) d_i^{(h)}(t,T)$, $d_i^{(h)}(t,T) \triangleq \frac{1}{\left|\mathbb{T}_h^{(t)}\right|} \sum_{s\in\mathbb{T}_h^{(t)}} |p_{i,s} - p_{i,t+T}|$, $h \in \{e, o\}$, and δ' is as defined in Theorem 3.

Proof See [14, Appendix B].

[1]Here, the dependence of the caching probability on t is omitted for brevity.

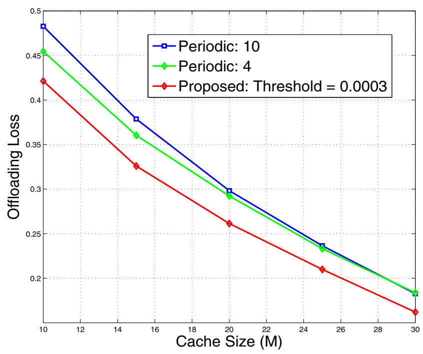

Fig. 2: Offloading loss as a function of the cache size.

V. Discussion

Theorem 4 suggests that the sBSs should update their caches at the time instants at which the error becomes large. The only relevant term is $\max\{\Delta_{t,T}^{(e)}, \Delta_{t,T}^{(o)}\} \leq \Delta_{t,T} \triangleq \frac{1}{\left|\mathbb{T}_e^{(t)} \bigcup \mathbb{T}_o^{(t)}\right|} \sup_{\Pi \in \mathcal{P}_\pi} \sum_{i=1}^{N} \sum_{s \in \mathbb{T}_o^{(t)} \bigcup \mathbb{T}_e^{(t)}} g(\pi_i) \left|p_{i,s} - p_{i,t+T}\right|$. The following cache update mechanism is employed:

1) Initialize $t = 0$ and $T = 0$. Update the caches randomly.
2) If $\Delta_{t,T} > \texttt{threshold}$, then update the caches using the caching probability obtained by solving $\hat{\Pi}_{t+T}^* = \arg\min_{\Pi^{(t+T)} \in \mathcal{P}_\pi} \mathcal{T}(\Pi^{(t+T)}, \hat{\mathcal{P}}^{(t+T-1)})$, where $\hat{\mathcal{P}}^{(t+T-1)}$ is the estimate obtained using (1), and set $T = t$. Here, $\texttt{threshold} > 0$ determines the error achieved.
3) Set $t \to t+1$ and go to step 2.

The behavior of the offloading loss for varying cache sizes is shown in Fig. 2. The setting comprises sBSs and users distributed according to a PPP with densities $\lambda_B = 0.00001$ and $\lambda_u = 0.0001$, respectively. The number of files is $N = 50$, and the coverage of the BS and sBSs are 1000 m and 500 m, respectively. We let $\gamma = 500$. The requests follow a Poisson arrivals with rate $\lambda_r = 0.05$. The requests for the files are generated using the Zipf distribution with the parameter chosen uniformly in the interval $(0.9, 1.1)$, and the file index is randomly and uniformly permuted in every time slot. This results in an independent but non-stationary arrival of requests. The requests from a typical user at the origin are used to evaluate the offloading loss. We assume perfect knowledge of $\Delta_{t,T}^{(h)}$ at the BS. Fig. 2 shows a plot of the offloading loss with $B = R_0$ as a function of the cache size for the following scenarios: (i) cache update mechanism with $\texttt{threshold} = 0.0003$, and (ii) periodic cache update with periods 10 and 4 slots, for cache sizes $10, 15, 20, 25$ and 30. In the periodic caching scheme, we follow the aforementioned caching policy; however, the caches are updated periodically. From Fig. 2, it is clear that the proposed cache update algorithm outperforms periodic caching, since the cache is updated only when required. Also, increasing the update period degrades the performance as the cache content items become less relevant.

The following remarks are in order (see (11)): (a) The error ϵ increases linearly with N. To compensate for larger values of N, the waiting time t should be of the order of N^2; a similar observation was also made in [9]. As λ_u increases, a lower value of δ can be achieved. In general, as $\lambda_u \to \infty$, $\delta = 0$ cannot be achieved due to the dependence of the stochastic process across time, *i.e.*, $\beta(a) > 0$, $a > 0$. (b) The error ϵ decreases as t increases. When the requests are i.i.d., $a_{\max} = 1$, and hence ϵ is small. Thus, when the requests are correlated we incur a penalty of $a_{\max}$, since the error decreases as $\sqrt{1/(t/a_{\max})}$ compared to $\sqrt{1/t}$ for i.i.d. requests. The error can be reduced by choosing $a_{\max} = 1$, *i.e.*, $a_i = 1$, $i = 1, \ldots, 2m$. Since $\beta(x)$ is a monotonically decreasing function of x, the probability of achieving a lower error is very small, indicating a tradeoff between the error and the probability with which the bound in (11) holds. Also, lower values of $\delta^{'}$ result in a higher error. (c) The error ϵ increases with $\frac{\alpha_{\max}}{\alpha_{\min}}$. The higher this ratio, the larger the variation in the number of requests. On the other hand, the lower this ratio, the lesser the error; which indicates a greater number of requests. The non-stationarity of the process is captured through $\Delta_{t,T}^{(h)}$, $h \in \{e, o\}$. For a stationary process $\Delta_{t,T}^{(h)} = 0$, $h \in \{e, o\}$. (d) When the user requests are i.i.d., the error does not vanish as $t \to \infty$, because the Rademacher complexity will not go to zero as $t \to \infty$. This indicates the difficulty in estimating the offloading loss, or equivalently the popularity profile, for a given caching policy. (e) The only term that depends on T is $\max\{\Delta_{t,T}^{(e)}, \Delta_{t,T}^{(o)}\}$. The frequency with which the cache update should be done depends on $\Delta_{t,T}^{(h)}$, $h \in \{e, o\}$. For instance, if $\Delta_{t,T}^{(h)}$, $h \in \{e, o\}$, is high, then the updates should be more frequent.

VI. Concluding remarks

A learning-theoretic analysis of content caching in heterogenous networks with non-stationary, statistically dependent and unknown popularity profiles has been considered. A PAC result on the offloading loss gap is presented in Theorem 4, based on the following caching algorithm: At every slot t, the BS computes an estimate of the Rademacher complexity and the discrepancy based on the available requests. The optimal caching policy is employed at the BS and the cache content items at the sBSs are updated only if the discrepancy in the popularity profile is larger than a pre-specified threshold (to be determined based on the error tolerance).

Appendix A
Proof of Theorem 3

Consider the following:

$$\begin{aligned} \mathcal{A}_T(X_{1,t}) &\overset{(a)}{\leq} \sup_{\Pi \in \mathcal{P}_\pi} \left| \frac{\left|\mathbb{T}_e^{(t)}\right|}{r_t} \sum_{i=1}^{N} g(\pi_i) \left(\hat{p}_{i,t}^e - p_{i,t+T}\right) \right| \\ &+ \sup_{\Pi \in \mathcal{P}_\pi} \left| \frac{\left|\mathbb{T}_o^{(t)}\right|}{r_t} \sum_{i=1}^{N} g(\pi_i) \left(\hat{p}_{i,t}^o - p_{i,t+T}\right) \right| \\ &\overset{(b)}{\leq} \frac{\left|\mathbb{T}_e^{(t)}\right|}{r_t} \mathcal{A}_T(X_{1,t}^e) + \frac{\left|\mathbb{T}_o^{(t)}\right|}{r_t} \mathcal{A}_T(X_{1,t}^o), \end{aligned} \quad (12)$$

where $\hat{p}_{i,t}^h \triangleq \frac{1}{\left|\mathbb{T}_h^{(t)}\right|}\sum_{s\in\mathbb{T}_h^{(t)}} \mathbf{1}\{X(s)=i\}$, $h\in\{e,o\}$, and $\mathcal{A}_T(X_{1,t}^{(h)}) \triangleq \sup_{\Pi\in\mathcal{P}_\pi}\left|\sum_{i=1}^N g(\pi_i)\left(\hat{p}_{i,t}^h - p_{i,t+T}\right)\right|$. In (12), (a) follows from the triangle inequality, and (b) follows from the convexity of $|\cdot|$ and the sup. We can further write

$$\Pr\{\mathcal{A}_T(X_{1,t})>\epsilon|n_u=j\} \leq \Pr\left\{\frac{\left|\mathbb{T}_e^{(t)}\right|}{r_t}\mathcal{A}_T^e(X_{1,t}) + \frac{\left|\mathbb{T}_o^{(t)}\right|}{r_t}\mathcal{A}_T^o(X_{1,t}) > \epsilon|n_u=j\right\} \quad (13)$$

$$\stackrel{(a)}{\leq} \Pr\{\mathcal{A}_T(X_{1,t}^e)>\epsilon|\mathcal{E}_j\} + \Pr\{\mathcal{A}_T(X_{1,t}^o)>\epsilon|\mathcal{E}_j\}, \quad (14)$$

where $\mathcal{E}_j \triangleq \{n_u=j\}$, and (a) follows from the union bound. We now bound the term corresponding to the even samples (the bound on the term corresponding to the odd samples follows similarly). We begin with $\Pr\{\mathcal{A}_T(X_{1,t}^e)>\epsilon|n_u=j\}=\mathbb{E}[\mathbf{1}\{\mathcal{A}_T(X_{1,t}^e)>\epsilon\}|n_u=j]$. Since the indicator function is bounded, using [13, Proposition 1], we have the following upper bound:

$$\mathbb{E}[\mathbf{1}\{\mathcal{A}_T(X_{1,t}^e)>\epsilon\}|\mathcal{E}_j] \leq \mathbb{E}[\mathbf{1}\{\mathcal{A}_T(\tilde{X}_{1,t}^e)>\epsilon\}|\mathcal{E}_j] + \sum_{i=2}^m \beta(a_{2i-1}),$$

$$= \Pr\{\mathcal{A}_T(\tilde{X}_{1,t}^e)>\epsilon|\mathcal{E}_j\} + \sum_{i=2}^m \beta(a_{2i-1}), \quad (15)$$

where $\tilde{X}_{1,t}^e$ is as defined in Section IV. Since the conditioning is on $\mathcal{E}_j$, the time slot difference between adjacent even/odd blocks is deterministic, and the β-mixing is not conditioned on the event. Similarly, it can be shown that

$$\mathbb{E}[\mathbf{1}\{\mathcal{A}_T(X_{1,t}^o)>\epsilon\}|n_u=j] \leq \tilde{\alpha}_{t,T,o}(j) + \sum_{j=1}^{m-1}\beta(a_{2j}), \quad (16)$$

where $\tilde{\alpha}_{t,T,h}(j) \triangleq \Pr\{\mathcal{A}_T(\tilde{X}_{1,t}^h)>\epsilon|n_u=j\}$, $h\in\{e,o\}$, and $\mathcal{A}_T(\tilde{X}_{1,t}^e)$ (resp. $\mathcal{A}_T(\tilde{X}_{1,t}^o)$) is obtained by replacing each block of data in $X_{1,t}^e$ (resp. $X_{1,t}^o$) by $\tilde{X}_{1,t}^e$ (resp. $\tilde{X}_{1,t}^o$) in the definition of $\mathcal{A}_T(X_{1,t}^e)$ (resp. $\mathcal{A}_T(X_{1,t}^o)$). Using (16) and (15) in (14), we get

$$\Pr\{\mathcal{A}_T(X_{1,t})>\epsilon|n_u=j\} \leq \sum_{h\in\{e,o\}}\tilde{\alpha}_{t,T,h}(j) + \sum_{j=2}^{2m-1}\beta(a_j). \quad (17)$$

Since each of the events above involves a sum of blocks of data that are independent, we employ McDiarmid's inequality to get the following result.

Theorem 5: For any $\max\{\mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^e)],\mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^o)]\}<\epsilon$, and $m>0$, the following bound holds for all $j\geq 1$:

$$\sum_{h\in\{e,o\}}\Pr\{\mathcal{A}_T(\tilde{X}_{1,t}^h)>\epsilon|n_u=j\} \leq \exp\left\{-2mg_N\right\}, \quad (18)$$

where $g_N \triangleq \frac{R_0^2 a_{\min}^2 \min\{\epsilon_e^2,\epsilon_o^2\}\alpha_{\min}^2}{a_{\max}^2 B^2 \alpha_{\max}^2 N^2}$, $a_{\max} \triangleq \max_{1\leq i\leq 2m} a_i$, $a_{\min} \triangleq \min_{1\leq i\leq 2m} a_i$, and $\epsilon_h \triangleq \epsilon - \mathbb{E}[\mathcal{A}_T(\tilde{X}_{1,t}^h)]$, $h\in\{e,o\}$.

Proof See [14, Appendix C].

The bound in (18) is independent of j. Substituting the bound (18) into (17), and using the result in (8), we get

$$\Pr\left\{\mathcal{A}_T(X_{1,t})>\epsilon\right\} \leq \exp\left\{-\lambda_u\pi R^2\right\} + G_m, \quad (19)$$

where $G_m \triangleq \exp\left\{-\psi m\right\} + \sum_{i=2}^{2m-1}\beta(a_i)$, $\psi \triangleq \frac{2R_0^2 a_{\min}^2 \min\{\epsilon_e^2,\epsilon_o^2\}\alpha_{\min}^2}{a_{\max}^2 B^2 \alpha_{\max}^2 N^2}$. We need $\Pr\left\{\mathcal{A}_T(X_{1,t})>\epsilon\right\} < \delta/2$, which implies that

$$\min\{\epsilon_e,\epsilon_o\} > \frac{N\alpha_{\max}Ba_{\max}}{\alpha_{\min}R_0 a_{\min}}\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{2m}}, \quad (20)$$

where $\delta' \triangleq \delta/2 - \exp\left\{-\lambda_u\pi R^2\right\} - \sum_{i=2}^{2m-1}\beta(a_i) > 0$. But, $\epsilon_h = \epsilon - \mathbb{E}\left[\mathcal{A}_T(\tilde{X}_{1,t}^h)\right]$, $h\in\{e,o\}$. Using this in (20) results in the following constraint: $\epsilon > \mathcal{E}_{t,T} + \frac{N\alpha_{\max}Ba_{\max}}{\alpha_{\min}R_0 a_{\min}}\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{2m}}$, where $\mathcal{E}_{t,T} \triangleq \min\left\{\mathbb{E}\left[\mathcal{A}_T(\tilde{X}_{1,t}^e)\right],\mathbb{E}\left[\mathcal{A}_T(\tilde{X}_{1,t}^o)\right]\right\}$. Using this constraint for ϵ, the bound in the theorem follows with a probability of at least $(1-\delta)$.

REFERENCES

[1] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[2] P. Mansourifard, N. Golrezaei, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[3] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: modeling and tradeoffs," *EURASIP J. Wireless Commun. Net.*, vol. 2015:41, Feb. 2015.

[4] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1222– 1234, May 2016.

[5] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 1897–1903.

[6] ——, "Multi-armed bandit optimization of cache content in wireless infostation networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 51–55.

[7] B. N. Bharath and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2015, pp. 1–6.

[8] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. Int. Symp. Model. Opt. Mobile, Ad Hoc Wireless Net.*, May 2015, pp. 161–166.

[9] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.

[10] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[11] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Stochastic geometry and architecture of communication networks," *J. Telecom. Syst.*, vol. 7, no. 1, pp. 209–227, 1997.

[12] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1461–1465.

[13] V. Kuznetsov and M. Mohri, "Generalization bounds for time series prediction with non-stationary processes," in *Algorithmic Learning Theory*. Springer, 2014, pp. 260–274.

[14] B. N. Bharath, K. G. Nagananda, D. Gündüz, and H. V. Poor, "Learning-based content caching with time-varying popularity profiles," August 2017, in preparation.

[15] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2012.